AD_____

Award Number: W81XWH-04-1-0475

TITLE: Development of a Computer-Aided Diagnosis System for Early Detection of Masses Using Retrospectively Detected Cancers on Prior Mammograms

PRINCIPAL INVESTIGATOR: Jun Wei, Ph.D.

CONTRACTING ORGANIZATION: University of Michigan
Ann Arbor, MI 48109-1274

REPORT DATE: June 2007

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
01-06-2007	Annual	1 Jun 2006 – 31 May 2007

4. TITLE AND SUBTITLE

Development of a Computer-Aided Diagnosis System for Early Detection of Masses Using Retrospectively Detected Cancers on Prior Mammograms

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-04-1-0475

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Jun Wei, Ph.D.

E-Mail: jvwei@umich.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

University of Michigan
Ann Arbor, MI 48109-1274

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES – Original contains colored plates: ALL DTIC reproductions will be in black and white.

14. ABSTRACT
The goal of this project is to develop a computer-aided diagnosis (CAD) system for mass detection using advanced computer vision techniques that will be trained with retrospectively detected cancers on prior mammograms. The new CAD system will be combined with our existing CAD system. When fully developed, the new dual CAD system should increase the sensitivity of detecting cancers at the early stage without compromising the sensitivity for other cancers.
During this project year, we have performed the following tasks: (1) continue to collect the data sets of digitized film mammograms for testing our CAD system, (2) investigation of a bilateral approach to reduce the false positives (FPs) on single CAD system, (3) develop image processing techniques for improvement of mass detection on prior mammograms, and (4) continue to develop a two-view information fusion method to improve the performance of single CAD system.
In summary, we have investigated a number of areas in CAD of mammographic masses and evaluated the new techniques for mass detection on mammograms. We have made progress in three of the tasks proposed in the project. We have found that our new computer-vision techniques can improve the performance of the CAD systems. We will continue the development of the CAD system in the coming years.

15. SUBJECT TERMS
Breast cancer, mammography, CAD

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			USAMRMC
U	U	U	UU	85	**19b. TELEPHONE NUMBER** *(include area code)*

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

Table of Contents

(4) Introduction

Recent clinical studies have proved that computer-aided diagnosis (CAD) systems are helpful for improving cancer detection by radiologists on mammograms[1-6]. To evaluate the effectiveness of a CAD system in detecting cancers that are likely to be missed by radiologists, one way is to study its accuracy in detecting missed cancers on prior mammograms (the mammograms in previous exams on which the cancer can be seen retrospectively). Several studies have demonstrated that CAD systems have potential ability to detect missed cancers on prior mammograms[7-11]. However, the performance of a CAD system on prior mammograms is generally much lower than their performance on the current mammograms (the mammogram on which cancer is detected). Recently, one study investigated the performance change between prior mammograms and current mammograms when using the CAD system trained by current mammograms and another by prior mammograms. It was concluded that CAD schemes trained with the current mammograms do not perform optimally in detecting masses depicted on prior images and vice versa.

The goal of this proposed project is to develop a CAD system using advanced computer vision techniques to detect masses using retrospectively detected cancers on prior mammograms and incorporate the developed CAD system into our current CAD system. We hypothesize that a dual CAD system, which combines a system trained with subtle lesions retrospectively seen on prior mammograms and a system trained with cancers detected on current mammograms, should increase the sensitivity of detecting cancers at the early stage without compromising its ability to detect less subtle cancers. To accomplish this goal, we will (1) collect a large database of masses on digitized prior and current film mammograms (DFMs) for training and testing the CAD system, (2) develop single-view computer vision techniques for mass detection and classification in prior DFMs, (3) reduce false positives (FPs) by correlation of image information from two-view mammograms, (4) combine the new CAD system with our current CAD system without an increase in overall FPs, and (5) perform ROC study to evaluate the effects of CAD on radiologists' accuracy in detecting subtle cancers. Although we do not plan to develop such a system for digital mammograms because there will not be enough prior digital mammograms with cancers available for the development, the general methodology developed in this study can be adapted to CAD systems for digital mammograms in the future.

At the conclusion of this project, we expect that a fully automated CAD system will be developed which can be used for detection of masses on DFMs. The general methodology developed in this study may also be adapted to develop similar software for other CAD systems. The significance of this project is that it will develop a CAD system which can further improve radiologists' accuracy in detecting breast cancers at an early stage. Since early detection and treatment can reduce breast cancer mortality rate, the CAD system will be useful for increasing the effectiveness of mammographic screening.

(5) Body

The current year (6/1/06-5/31/07) is the third year of the project. We have requested and obtained approval for a no cost time extension of the project so that this is a regular annual progress report instead of a final report. We will describe in the following details of the studies that we performed this year.

(A) Collection of a Database of Digitized Screen-film Mammograms (DFM) with Multiple Examinations

In this project year, we continue to collect a data set of digitized screen-film mammogram from patient files in the Department of Radiology at the University of Michigan with Institutional Review Board (IRB) approval. Two independent data sets of mammograms were collected for this study; one contained mammograms with masses and the other contained normal mammograms. The normal data set was used to estimate the false positive (FP) marker rates during testing[12-14]. To date, the mass data set contained 220 cases with 220 masses. 190 cases included the current mammograms on which the mass was detected by radiologists, and the prior mammograms obtained from previous exams. 30 cases only had the current mammograms. In total, 886 mammograms including 440 current mammograms and 446 prior mammograms were collected. The true location of each mass was identified by an experienced Mammography Quality Standards Act (MQSA) radiologist. The radiologist also measured the mass size and provided descriptions of the mass margin, shape, conspicuity, and breast density.

(B) Investigation of a Bilateral Approach to Reduce the FPs on Single CAD System

In an effort to improve the performance of our single CAD system, the first study of this year is to investigate an FP reduction method based on analysis of bilateral mammograms for computerized mass detection systems. Our recent paper has been accepted for publication on the Medical Physics Journal[15]. The study is summarized in the following.

1. Data Set

A database of mammograms was collected from patient files at the Department of Radiology with Institutional Review Board (IRB) approval. Two data sets are used: a mass data set containing bilateral digitized mammograms with malignant or benign masses and a no-mass data set containing bilateral digitized mammograms without masses, verified by an experienced radiologist. All cases had four mammographic views, the CC view and the MLO view mammogram for both breasts. The mass set contained 276 cases so that 552 bilateral pairs were available. The no-mass data set contained 65 cases so that 130 bilateral pairs were available. Fifty cases of the no-mass set were consecutive normal screening cases from our patient files with an additional 15 cases visually judged by radiologists to be dense breasts. The mass data set was used to estimate the detection sensitivity and the no-mass data set was used for estimating the FP rate. In the mass data set, each patient had a biopsy-proven mass in one of the breasts, resulting in a total of 276 masses, 166 of which were benign and 110 malignant. An MQSA radiologist identified the location of the masses, measured the mass sizes as the longest

dimension seen on the two-view mammograms, provided descriptors of the mass shapes and mass margins, and also provided an estimate of the breast density in term of BI-RADS category.

2. Methods

In order to improve the performance of our CAD system, we developed a new bilateral CAD system that combines the unilateral features with the bilateral features to reduce FPs. Our bilateral CAD system consists of five steps: (1) mass candidate (MC) detection, (2) corresponding ROIs (CR) extraction, (3) feature analysis, (4) feature combination, and (5) bilateral CAD system generation. Figure 1 shows the block diagram for our bilateral CAD system.

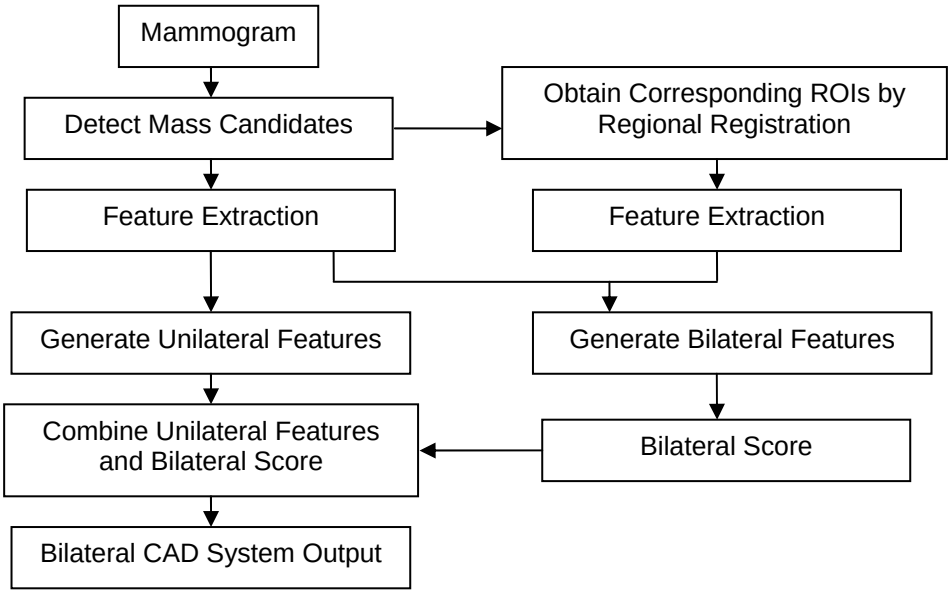


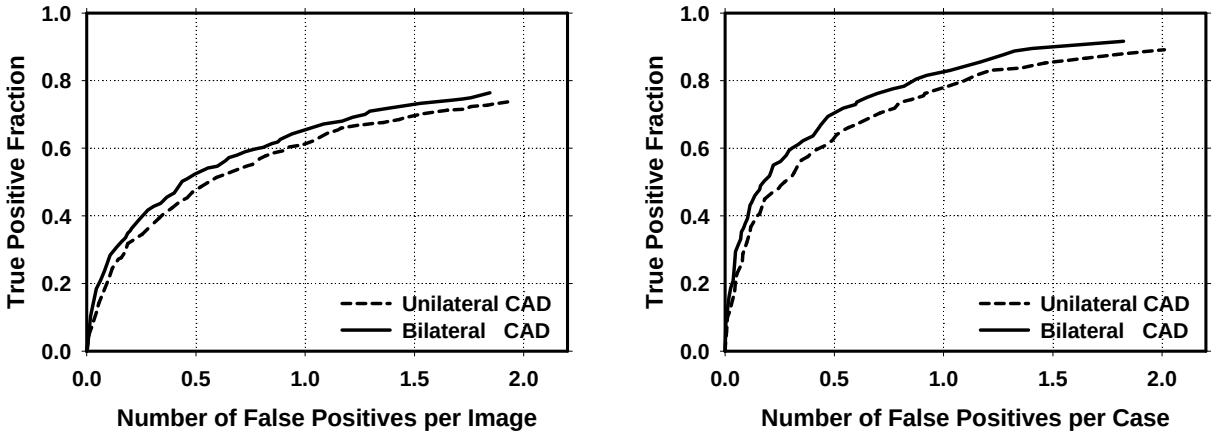
Figure 1. The block diagram of the bilateral CAD system for FP reduction on mammograms.

The mass candidates on the individual mammograms are detected by the prescreening step of our unilateral CAD system[16,17]. A gradient field analysis is applied to the mammogram and the locations of high gradient convergence are identified as mass candidates. An ROI of 256×256 pixels is then centered at each location of high gradient convergence. For each candidate, regional registration technique[16-19] is used to define an ROI that is "symmetrical" to the object location on the contralateral mammogram. An ROI of 256×256 pixels is then centered at the triangle as the contralateral ROI. For the feature analysis, SGLD texture features and morphological features are extracted from both the ROIs containing the detected mass candidate and its contralateral ROI. Let $MC[i, j]$ and $CR[i, j]$ be the i^{th} feature of the j^{th} mass candidate and the i^{th} feature of the j^{th} corresponding ROI, respectively. The i^{th} bilateral feature (BF) is derived from the i^{th} unilateral feature by the expression below.

$$BF(i,j) = \frac{Max(MC[i,j], CR[i,j])}{Min(MC[i,j], CR[i,j])} \quad (1)$$

Using the bilateral features as the input predictor variables, a linear discriminant analysis (LDA) classifier is trained to merge the features into a bilateral score. The LDA is trained with a leave-one-case-out resampling scheme for stepwise feature selection within the training set. The bilateral score incorporates the "symmetry" information to differentiate symmetric (likely FPs) and asymmetric (likely masses) tissues on the left and right breasts. To merge the information from the unilateral and the bilateral features in our CAD system, a new feature space is formed by combining the unilateral features and the bilateral score of the mass candidate. Finally, the bilateral CAD system output score is obtained by a second LDA classifier that is trained to differentiate the true mass and FPs in the new feature space, again using the training set.

3. Results

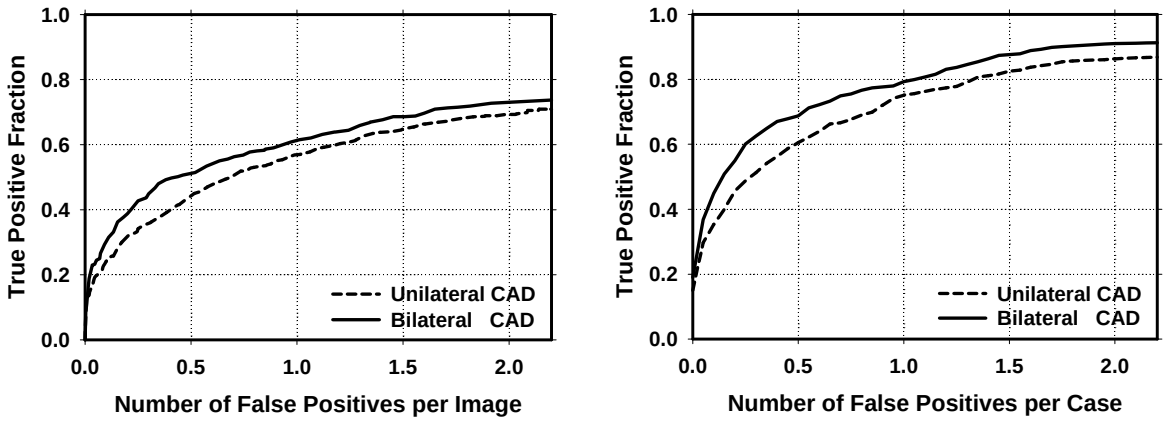


(a) (b)

Figure 2. (a) Image-based and (b) case-based average test FROC curves from the unilateral and the bilateral CAD systems. The FP rates were estimated from detection on mammograms in the test subsets with masses.

Figure 2 shows the average free response receiver operating characteristic (FROC) curves for the test sets using the unilateral and bilateral CAD systems. The the corresponding trained LDA classifiers have been used for FP reduction and the FP rates were estimated from the test subsets with masses. The bilateral CAD system achieved a case-based sensitivity of 70%, 80%, and 85% at average FP rates of 0.53, 0.87, and 1.15 FPs/image, respectively, on the test data set. In comparison to the average FP rates for the unilateral CAD system of 0.70, 1.11, and 1.46 FPs/image, respectively, at the corresponding sensitivities, the FP rates were reduced by 24%, 21%, and 21% with the bilateral symmetry information. Figure 3 shows the average test FROC curves for the unilateral and bilateral CAD systems with the FP rates estimated on the set of no-mass mammograms. Figure 4 compares the average test FROC curves for the unilateral and bilateral CAD systems on malignant cases only. Figure 5 shows the average test FROC curves

for the unilateral and bilateral CAD systems with the sensitivities estimated on malignant cases only and the FP rates estimated on the set of no-mass mammograms. We employed the jackknife FROC (JAFROC) analysis[23,24] to evaluate the difference in the FROC curves obtained from the unilateral CAD system and the bilateral CAD system. The difference between the figures of merit (FOMs) for the unilateral and the bilateral CAD systems was statistically significant ($p<0.05$).

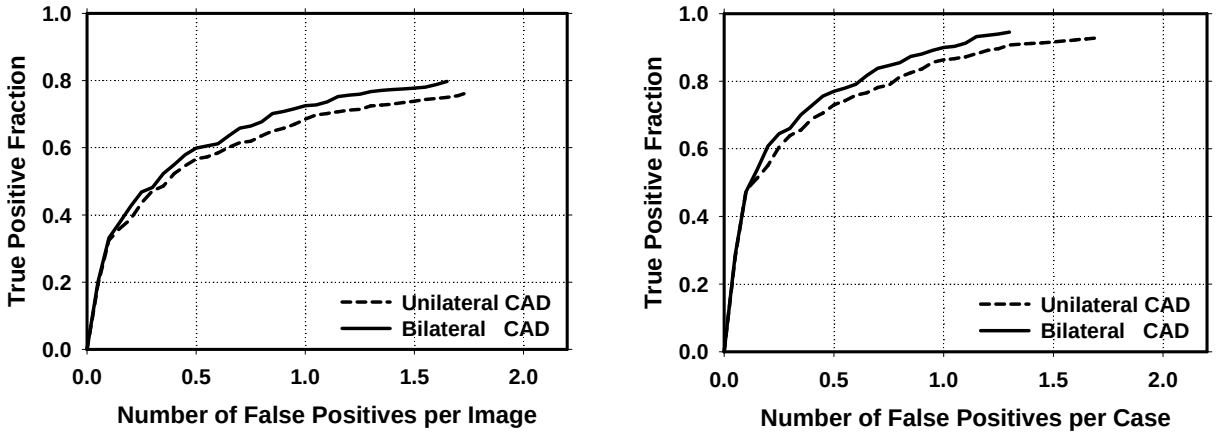


(a) (b)

Figure 3. (a) Image-based and (b) case-based average test FROC curves from the unilateral and the bilateral CAD systems. The FP rates were estimated from detection on mammograms in the no-mass data set.

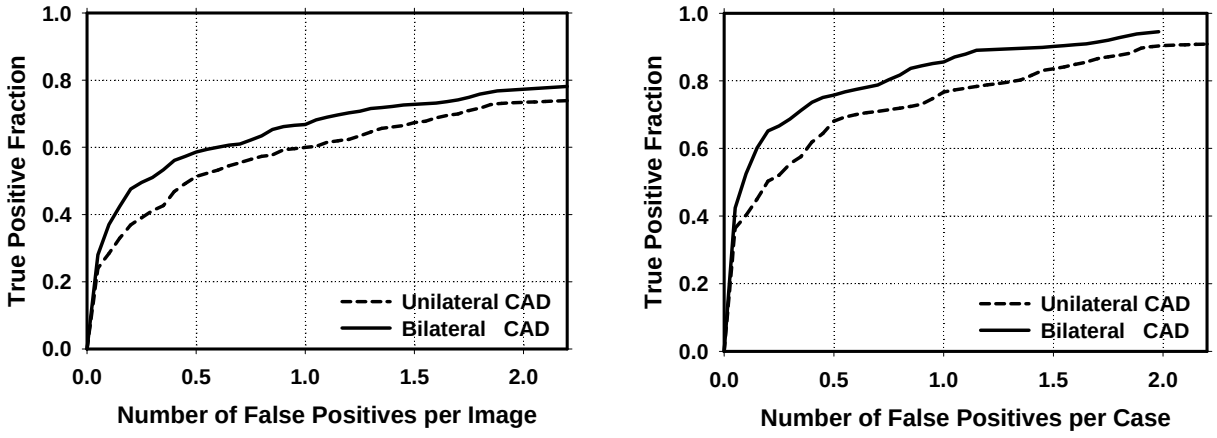
4. Discussion and Conclusion

Symmetry between breast structures in bilateral pairs of mammograms is an important feature used by radiologists for mass detection or FP reduction. Similar structures that appear in both the right and left mammograms are more likely to be normal tissue than abnormal lesions. Our bilateral analysis translates this human intelligence to computer vision that can recognize the symmetry of breast tissue on bilateral mammograms to improve detection accuracy. To our knowledge, this FP reduction strategy for mass detection has not been reported previously. Our results demonstrate that the bilateral features can be utilized to differentiate the similarity and dissimilarity between tissues at corresponding locations in the bilateral views, and can be useful for improving the performance of a unilateral CAD system by further reducing the FPs.



(a) (b)

Figure 4. (a) Image-based and (b) case-based average test FROC curves from the unilateral and bilateral CAD systems for detection on cases with malignant masses only. The FP rates were estimated from the test subset with masses.



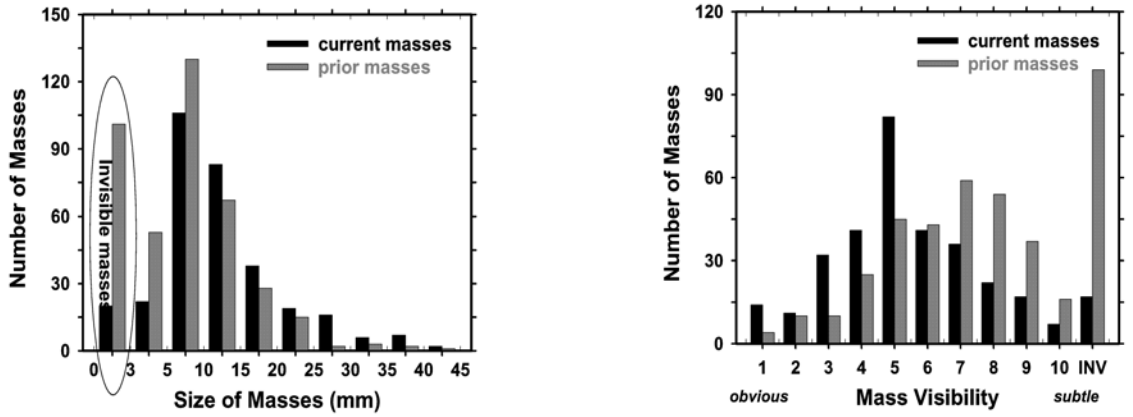
(a) (b)

Figure 5. (a) Image-based and (b) case-based average test FROC curves from the unilateral and bilateral CAD systems for detection on cases with malignant masses only. The FP rates were estimated from the no-mass data set.

(C) Development of Image Processing Techniques for Improvement of Mass Detection on Prior Mammograms

The second study of this year is to develop image processing techniques for improvement of CAD on prior mammograms. Our results were presented at the SPIE meeting in 2007[20]. Also, these techniques have been applied to mass detection on full field digital mammograms (FFDM) and screening film mammograms (SFM). We found that they were useful for improving the accuracy of mass detection on current mammograms as well. A journal article with the results had been published in Academic radiology 2007[21]. The study is summarized in the following.

1. Data Set



(a) mass sizes (b) mass visibility

Figure 6. Histogram of the sizes and visibility for 299 masses on current mammograms and 301 masses on priors in our data set. The size of the masses in this data set ranged from 3 to 42 mm. The visibility is evaluated on a 10-point rating scale with 1 representing the most visible masses and 10 the most difficult case relative to the cases seen in their clinical practice. The masses that were not visible were plotted in the column labeled as "INV".

All mammograms in this study were collected from patient files in the Department of Radiology at the University of Michigan with Institutional Review Board (IRB) approval. The mammograms were digitized with a LUMISYS 85 laser film scanner with a pixel size of 50μm×50μm and 4096 gray levels. The full resolution mammograms were first smoothed with a 2×2 box filter and subsampled by a factor of 2, resulting in images with a pixel size of 100μm×100μm. These images were used for input to our CAD system. The data set we used in this study contained 159 cases. Each exam had two mammographic views, resulting in a total of 318 current mammograms and 402 prior mammograms. Forty-two patients had two years of prior examinations. All mammograms were obtained before biopsy. There were 159 biopsy-proven masses in this data set. Figures 6 showed the histograms of mass sizes and visibility, respectively, for the comparison of current and prior masses. The size of a mass was estimated

as its longest diameter seen on the mammograms. The visibility of the masses was rated by an experienced radiologist on a 10-point scale with 1 representing the most visible masses and 10 the most difficult case relative to the cases seen in clinical practice. The mass size ranged from 3 to 42 mm (mean size: 14.3±8.6 mm on current mammograms and 10.9±6.6 mm on prior mammograms) and the visibility ratings extended over the entire range. For the current mammograms, 140 of the masses were visible on both views and 19 visible on only one view. For the prior mammograms, 100 masses were visible on both views and 101 visible only on one view. Therefore, there were 299 visible and 19 invisible masses on current mammograms and 301 visible and 101 invisible masses on prior mammograms if the masses were counted independently by mammographic view.

2. Methods

Our CAD system consists of five processing steps: 1) pre-screening of mass candidates, 2) identification of suspicious objects, 3) extraction of morphological and texture features, and 4) classification between the normal and the abnormal regions by using rule-based and LDA classifiers.

For the pre-screening stage, we developed a new prescreening technique in which gradient field analysis was combined with Hessian analysis to identify mass candidates. Both gradient field and Hessian analyses were designed to enhance approximately circular structures on mammograms and to suppress the objects with other shapes. Gradient field analysis used the information of gradient field directions and Hessian analysis used the second derivatives by solving for the eigenvalues of the Hessian matrix. After this enhancement filtering, the local maxima within the breast region were identified as the mass candidates on each mammogram. The suspicious structure in each identified location was initially extracted by a seed-based region growing method. An active contour method was then used to further refine the initial segmentation. Morphological, gray level histogram and run-length statistics (RLS) features were extracted from the original region of interest (ROI) and the orientation field of the ROI for reduction of FPs.

The hold-out method was used for training and testing our CAD system. We randomly separated the entire data set by case into two independent subsets, the training subset including 78 cases with 156 current and 200 prior mammograms and the test subset including 81 cases with 162 current and 202 prior mammograms. The training included selection of proper parameters and features for the classifier in the CAD system. Once the training was completed, the parameters and features were fixed for testing. The new system was trained by using prior mammograms in the training set only. The performance of the new system was compared with that of the previous CAD system on the current and prior mammograms in the test set.

During training, feature selection with stepwise LDA was employed to obtain the best feature subset and reduce the dimensionality of the feature space to design an effective classifier. The detailed procedure has been described elsewhere[17]. Briefly, at each step one feature was entered or removed from the feature pool by analyzing its effect on the selection criterion, which was chosen to be the Wilks' lambda in this study. Since the appropriate threshold values for feature entry, feature elimination, and tolerance of feature correlation were unknown, we used an automated simplex optimization method to search for the best combination of thresholds in the parameter space. The simplex algorithm used a leave-one-case-out resampling method within

the training subset to select features and estimate the weights for the LDA classifier. To have a figure-of-merit to guide feature selection, the test discriminant scores from the left-out cases were analyzed using receiver operating characteristic (ROC) methodology. The accuracy for classification of masses and FPs was evaluated as the area under the ROC curve, A_z. In this approach, feature selection was performed without the left-out case so that the test performance would be less optimistically biased. However, the selected feature set in each leave-one-case-out cycle could be slightly different because every cycle had one training case different from the other cycles. In order to obtain a single trained classifier to apply to the hold-out test subset, a final stepwise feature selection was performed with the best combination of thresholds, found in the simplex optimization procedure, on the entire training subset to obtain the final set of features and estimate the weights of the LDA. Note that the entire process of feature selection and classifier weight estimation was performed within the training subset. The LDA classifier with the selected feature set was then fixed and applied to the test subset.

3. Results

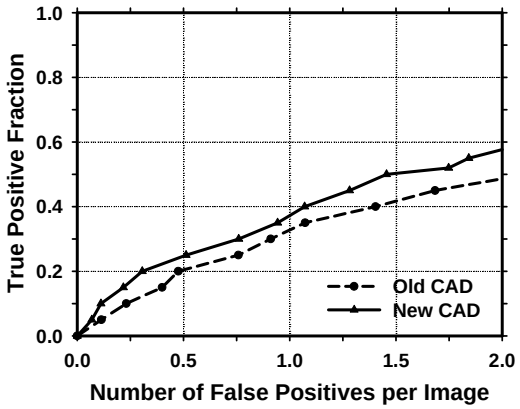


Figure 7. Image-based test FROC curves on prior mammograms. Old CAD: detection by the previous CAD system trained on both current and prior mammograms. New CAD: detection by the CAD system trained on prior mammograms.

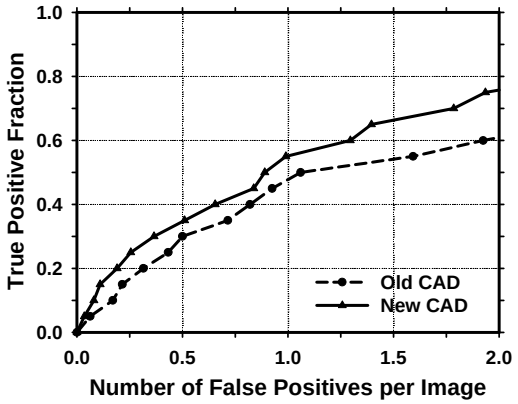


Figure 8. Case-based test FROC curves on prior mammograms. Old CAD: detection by the previous CAD system trained on both current and prior mammograms. New CAD: detection by the CAD system trained on prior mammograms.

Figures 7 and 8 showed the image-based and case-based FROC curves for detection of masses on prior mammograms, respectively. The case-based sensitivities for detection of masses on the prior mammograms (typically subtle masses) in the test subset were 56%, and 35% at 1 and 0.5 FPs/image by using the new CAD system in comparison to 48%, and 32% at the same FP rates by using the previous CAD system. The improvement with the new system on prior mammograms was statistically significant ($p = 0.036$). When the new system was applied to the detection of masses on the current mammograms (typically average masses) in the test subset, the case-based sensitivities were 77% and 70% at 1 and 0.5 FPs/image in comparison to 75% and 56% at the same FP rates by using the previous CAD system. The difference in the two FROC curves for detection of average masses on current mammograms was not statistically

different ($p = 0.184$) by JAFROC analysis. Image-based and case-based FROC curves for detection of masses on current mammograms were shown in Figures 9 and 10, respectively.

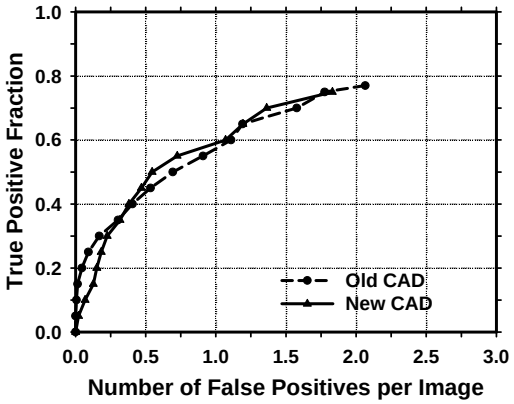


Figure 9. Image-based test FROC curves on current mammograms. Old CAD: detection by the previous CAD system trained on both current and prior mammograms. New CAD: detection by the CAD system trained on prior mammograms.

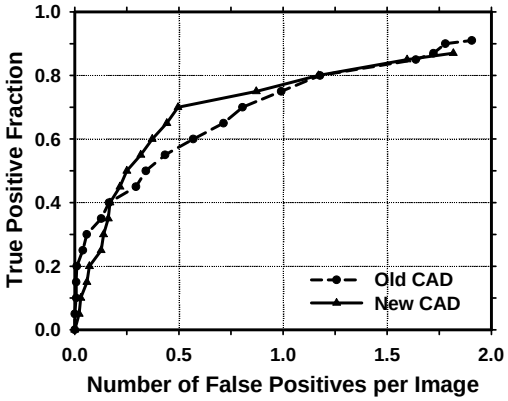


Figure 10. Case-based test FROC curves on current mammograms. Old CAD: detection by the previous CAD system trained on both current and prior mammograms. New CAD: detection by the CAD system trained on prior mammograms.

4. Discussion and Conclusion

In this study, we improved the accuracy of a CAD system for detection of subtle masses on prior mammograms. A new prescreening method was developed to improve the sensitivity of mass detection. A new mass segmentation method that combined a seed-based region growing method with active contour method was also designed. RLS features were extracted from the original ROIs and the newly derived orientation field of the ROIs for FPs reduction. Our CAD system can significantly improve the performance of mass detection on prior mammograms without a trade-off in the detection of masses on current mammograms. It is expected that the new CAD system can increase the overall accuracy for detection of subtle early-stage breast cancers.

(D) Continuation of Development of a Two-view Information Fusion Method to Improve the Performance of Single CAD System

The third study performed in this project year is to continue to develop a two-view information fusion method to improve the performance of our CAD system for mass detection. Our results were presented at the RSNA in Chicago[22] in November of 2006. We have made good progress on this part and are preparing a journal paper. The study is summarized in the following.

1. Data Set

All mammograms in this study were collected from patient files in the Department of Radiology at the University of Michigan with Institutional Review Board (IRB) approval. In this study, two data sets were collected: a mass set with biopsy-proven unilateral malignant or benign masses and a normal set containing bilateral mammograms. The mass set contained 469 cases with 469 biopsy proven masses, of which 190 were malignant and 279 benign. Each case contained two mammographic views (CC view and MLO view or the lateral view). The normal set was consisted of 50 consecutive normal screening cases from our patient files and an additional 15 cases visually judged by radiologists to be dense breasts. Each normal case contained 4 mammographic views from bilateral breasts. The normal data set was only used for estimating the FP rate during testing. An experienced MQSA radiologist identified the locations of masses by examining all available information including the diagnostic mammograms and reports. In these 469 mass cases, 19 masses (4%) can be seen only on one mammographic view.

2. Methods

In order to improve the overall performance of our CAD system for detection of masses, we developed a two-view fusion technique which combines the information from two mammographic views. Our method in this study is based on two assumptions: (1) the corresponding true masses on two different mammographic views will exhibit higher similarity than the FPs detected by the CAD system, (2) the morphological and texture features of the same mass on different views will also show similar properties and mass pairs (TP-TP pairs) can be distinguished from false pairs (TP-FP, FP-TP, FP-FP pairs) in the combined feature space from two different mammographic views. A schematic of our two-view system is shown in Figure 11.

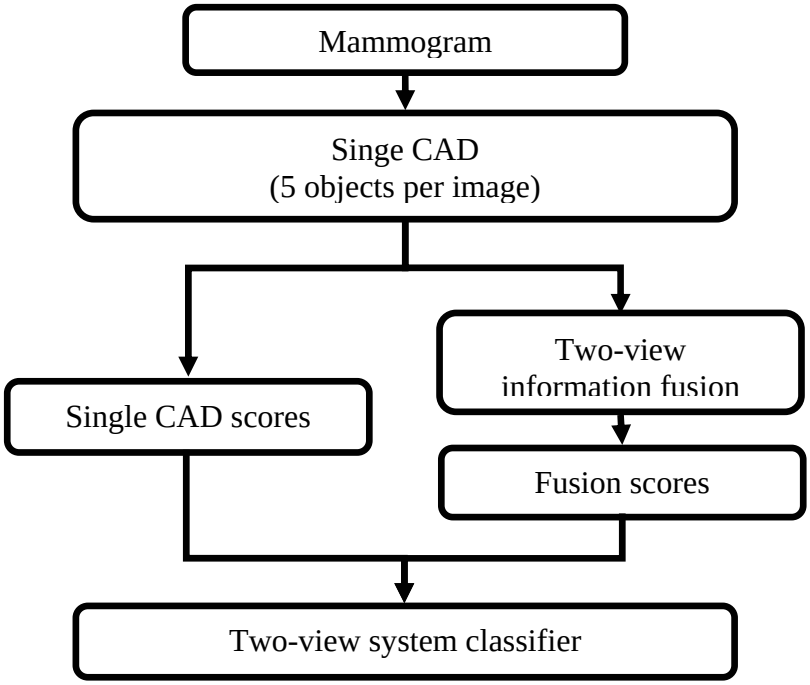


Figure 11. Schematic diagram of our two-view CAD system for mass detection on mammograms. The system is developed for screening mammography in which all masses, regardless of malignant or benign, are considered positive.

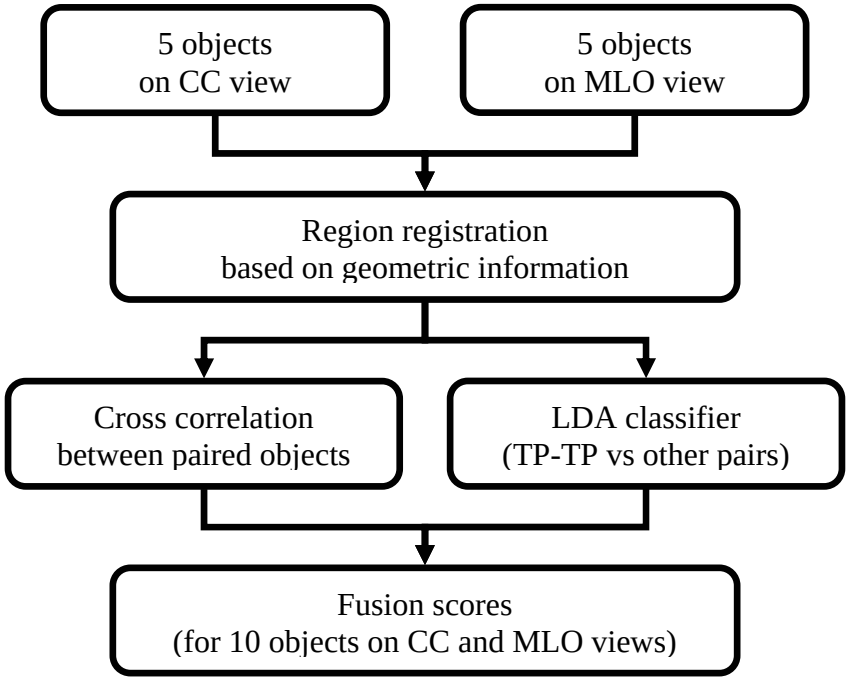
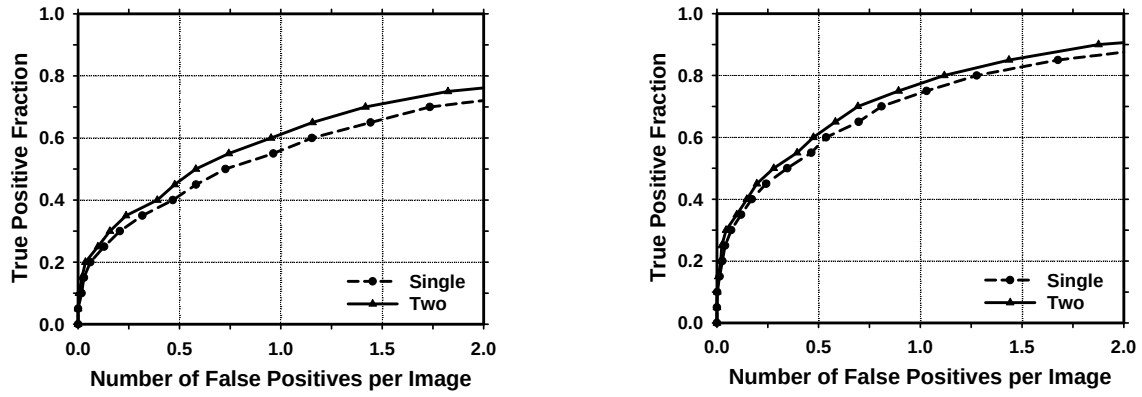


Figure 12. Block diagram of the two-view information fusion for suspicious objects on CC and MLO views of the same breast.

The key process of our two-view CAD system is the information fusion step in which the potential similar suspicious objects on different mammographic views are paired together and a classifier merges the two-view information and provides a unique fusion score for each individual suspicious object. For a deformable object like the breast under compression, the corresponding locations of an object in the two views cannot be determined exactly based on the two projection mammograms. Our two-view information fusion scheme consists of four steps: (1) region registration by using geometric information, (2) similarity measure between paired objects, (3) classification of TP-TP pairs from other FP pairs, and (4) generalization of fusion score. Figure 12 shows the block diagram for two-view information fusion for suspicious objects on CC and MLO views of the same breast. The JAFROC method was used to compare the performance of our two-view CAD system with that of single CAD system statistically.

3. Results

We randomly separated the cases in our mass data set into two independent data sets: 238 cases with 472 images and 231 cases with 462 images. The training and testing were performed using the 2-fold cross validation method. The detection performance of the CAD system was assessed by FROC analysis. FP rate was estimated from the mammograms without masses. FROC curves were presented on a per-mammogram and a per-case basis. To evaluate the overall test performance, an average test FROC curve was obtained as described above. Figures 13(a) and 13(b) show the comparison of the test performance of the single-view CAD system and the two-view CAD systems by using image-based and case-based average FROC curves, respectively. The single-view CAD system achieved an FP rate of 2.3, 1.7, and 1.3 FPs/image at the case-based test sensitivities of 90%, 85% and 80%, respectively. With the two-view fusion system, the FP rates were reduced to 1.9, 1.4, and 1.1 FPs/image at the corresponding sensitivities, respectively. The improvement was found to be statistically significant (p<0.05) by the JAFROC method.



(a) (b)

Figure 13. (a) Image-based and (b) case-based average test FROC curves from the single-view and the two-view CAD systems. The FP rates were estimated from detection on mammograms in the no-mass data set.

4. Discussion and Conclusion

In this study, a two-view information fusion method was developed to improve the performance of our CAD system for mass detection on mammograms. The two-view CAD system is different from case-based scoring, in which detection of the same mass in either the CC view or the MLO view will be counted as a true positive, in that the detected objects in the two views are correlated and analyzed for similarity and the likelihood score of a mass detected in both views may be enhanced compared with FPs. Our results indicate that two-view fusion significantly improved the overall performance of the single-view CAD system.

(6) Key Research Accomplishments

- Continue to collect the data sets of digitized film mammograms with multiple examinations. (Task 1).
- Investigation of a bilateral approach to reduce the FPs on single CAD system (Task 2).
- Development of image processing techniques for improvement of mass detection on prior mammograms (Task 2).
- Continue to develop a two-view information fusion method (Task 3).

(7) Reportable Outcomes

As a result of the support by the USAMRMC BCRP grant, we have conducted studies to develop a computer-aided diagnosis system for early detection of masses using retrospectively detected cancers on prior mammograms. We have presented the results of these investigations in this project year and a journal article which was accepted for publication last year had been published in this project year. Also, we have one journal paper published in Academic Radiology and one journal paper accepted for publication in Medical Physics.

Journal Articles:

1. Wei J, Chan HP, Sahiner B, Hadjiiski LM, Helvie MA, Roubidoux MA, Zhou C, Ge J, "Dual system approach to computer-aided detection of breast masses on mammograms", Medical Physics, Vol. 33, No. 11, pp. 4157–4168, 2006.

2. Wei J, Hadjiiski LM, Sahiner B, Chan HP, Ge J, Roubidoux MA, Helvie MA, Zhou C, Wu YT, Paramagul C, Zhang Y, "Computer Aided Detection Systems for Breast Masses: Comparison of Performances on Full-Field Digital Mammograms and Digitized Screen-film Mammograms", Academic Radiology, Vol. 14, No. 6, pp. 659-669, 2007.

3. Wu YT, Wei J, Hadjiiski LM, Sahiner B, Zhou C, Ge J, Shi J, Zhang Y, Chan HP, "Bilateral analysis based false positive reduction for computer-aided mass detection", Medical Physics (in press).

Conference Proceeding:

1. Wei J, Sahiner B, Chan HP, Hadjiiski LM, Roubidoux MA, Helvie MA, Ge J, Zhou C, and Wu YT, "Computer-aided detection of breast masses on prior mammograms", Proc. SPIE, Vol. 6514, pp. 51-57, 2007.

Conference Presentation:

1. J. Wei, B. Sahiner, HP. Chan HP, M. A. Roubidoux, M. A. Helvie, YT. Wu, L. M. Hadjiiski, J. Ge, C. Zhou, "Computer-aided detection of breast masses on mammograms: Performance improvement using two-view information", Presentation at the 92nd Scientific Assembly and Annual Meeting of the Radiological Society of North America, Chicago, IL. November 26-December 1, 2006.

(8) Conclusions

During this project year, we first investigated an FP reduction method based on analysis of bilateral mammograms for computerized mass detection systems. Our results demonstrate that the bilateral features can be utilized to differentiate the similarity and dissimilarity between tissues at corresponding locations in the bilateral views, and can be useful for improving the performance of a unilateral CAD system by further reducing the FPs.

In a second study, we developed several image processing techniques for improvement of mass detection on prior mammograms. The new techniques can significantly improve the performance of mass detection on prior mammograms without a trade-off in the detection of masses on current mammograms. It is expected that the improved CAD system can increase the overall accuracy for detection of subtle early-stage breast cancers.

The third study performed in this project year is to continue to develop a two-view information fusion method to improve the performance of our CAD system for mass detection. The two-view CAD system is different from case-based scoring in that the detected objects in the two views are correlated and analyzed for similarity and the likelihood score of a mass detected in both views may be enhanced compared with FPs. Our results indicate that two-view fusion can significantly improve the overall performance of the single-view CAD system.

From the results of these studies, we found that our single CAD system can be improved by the new image processing techniques and the bilateral and two-view analyses. We have already shown in the previous annual progress reports that our proposed dual CAD system

approach is a very promising method to improve detection of subtle early breast cancers. In the coming project year, we plan to investigate the combination of the techniques developed in this project year with the dual CAD system approach. We expect that the dual CAD system will be further improved when combined with the new techniques.

(9) References

[1]T. W. Freer and M. J. Ulissey, "Screening mammography with computer-aided detection: Prospective study of 12,860 patients in a community breast center," Radiology 220, 781-786 (2001).

[2]M. A. Helvie, L. M. Hadjiiski, E. Makariou, H. P. Chan, N. Petrick, B. Sahiner, S. C. B. Lo, M. Freedman, D. Adler, J. Bailey, et al., "Sensitivity of noncommercial computer-aided detection system for mammographic breast cancer detection - A pilot clinical trial," Radiology 231, 208-214 (2004).

[3]R. L. Birdwell, P. Bandodkar, and D. M. Ikeda, "Computer-aided Detection with Screening Mammography in a University Hospital Setting," Radiology 236, 451-457 (2005).

[4]D. Gur, J. H. Sumkin, H. E. Rockette, M. A. Ganott, C. Hakim, L. A. Hardesty, W. R. Poller, R. Shah, and L. Wallace, "Changes in breast cancer detection and mammography recall rates after the introduction of a computer-aided detection system," J National Cancer Institute 96, 185-190 (2004).

[5]S. A. Feig, E. A. Sickles, W. P. Evans, and M. N. Linver, "Re. Changes in breast cancer detection and mammography recall rates after the introduction of a computer-aided detection system," J National Cancer Institute 96, 1260-1261 (2004).

[6]T. E. Cupples, "Impact of computer-aided detection (CAD) in a regional screening mammography program," Radiology 221(P), 520 (2001).

[7]N. Petrick, H. P. Chan, B. Sahiner, M. A. Helvie, and S. Paquerault, "Evaluation of an automated computer-aided diagnosis system for the detection of masses on prior mammograms," San Diego, 2000.

[8]G. M. Te Brake, N. Karssemeijer, and J. Hendriks, "Automated detection of breast carcinomas not detected in a screening program," Radiology 207, 465-471 (1998).

[9]D. M. Ikeda, R. L. Birdwell, K. F. O'Shaughnessy, E. A. Sickles, and R. J. Brenner, "Computer-aided detection output on 172 subtle findings on normal mammograms previously obtained in women with breast cancer detected at follow-up screening mammography," Radiology 230, 811-819 (2004).

[10]B. Zheng, W. F. Good, D. R. Armfield, C. Cohen, T. Hertzberg, J. H. Sumkin, and D. Gur, "Performance change of mammographic CAD schemes optimized with most-recent and prior image databases," Academic Radiology 10, 283-288 (2003).

[11]D. Gur, J. S. Stalder, L. A. Hardesty, B. Zheng, J. H. Sumkin, D. M. Chough, B. E. Shindel, and H. E. Rockette, "Computer-aided Detection Performance in Mammographic Examination of Masses: Assessment," Radiology 233, 418-423 (2004).

[12]K. F. O'Shaughnessy, R. A. Castellino, S. L. Muller, and K. Benali, "Computer-aided detection (CAD) on 90 biopsy-proven breast cancer cases acquired on a full-field digital mammography (FFDM) system," Radiology 221(P), 471 (2001).

[13]L. J. Warren Burhenne, S. A. Wood, C. J. D'Orsi, S. A. Feig, D. B. Kopans, K. F. O'Shaughnessy, E. A. Sickles, L. Tabar, C. J. Vyborny, and R. A. Castellino, "Potential contribution of computer-aided detection to the sensitivity of screening mammography," Radiology 215, 554-562 (2000).

[14]R. E. Brem, J. W. Hoffmeister, J. A. Rapelyea, G. Zisman, K. Mohtashemi, G. Jindal, M. P. DiSimio, and S. K. Rogers, "Impact of breast density on computer-aided detection for breast cancer," AJR Am J Roentgenol. 184, 439-444 (2005).

[15]Y.-T. Wu, J. Wei, L. M. Hadjiiski, B. Sahiner, C. Zhou, J. Ge, J. Shi, Y. Zhang, and H.-P. Chan, "Bilateral analysis based false positive reduction for computer-aided mass detection," Medical Physics (in press) (2007).

[16]J. Wei, B. Sahiner, L. M. Hadjiiski, H. P. Chan, N. Petrick, M. A. Helvie, C. Zhou, and Z. Ge, "Computer aided detection of breast masses on full-field digital mammograms: false positive reduction using gradient field analysis," Proc. SPIE Medical Imaging 5370, 992-998 (2004).

[17]J. Wei, B. Sahiner, L. M. Hadjiiski, H. P. Chan, N. Petrick, M. A. Helvie, M. A. Roubidoux, J. Ge, and C. Zhou, "Computer aided detection of breast masses on full field digital mammograms," Medical Physics 32, 2827-2838 (2005).

[18]L. M. Hadjiiski, H. P. Chan, B. Sahiner, N. Petrick, and M. A. Helvie, "Automated registration of breast lesions in temporal pairs of mammograms for interval change analysis - local affine transformation for improved localization," Medical Physics 28, 1070-1079 (2001).

[19]C. Zhou, H. P. Chan, C. Paramagul, M. A. Roubidoux, B. Sahiner, L. M. Hadjiiski, and N. Petrick, "Computerized nipple identification for multiple image analysis in computer-aided diagnosis," Medical Physics 31, 2871-2882 (2004).

[20]J. Wei, B. Sahiner, H.-P. Chan, L. M. Hadjiiski, M. A. Roubidoux, M. A. Helvie, J. Ge, C. Zhou, and Y.-T. Wu, "Computer-aided detection of breast masses on prior mammograms," Proc. SPIE, Vol. 6514, pp. 51-57, 2007.

[21]J. Wei, L. M. Hadjiiski, B. Sahiner, H. Chan, J. Ge, M. A. Roubidoux, M. A. Helvie, C. Zhou, Y. Wu, C. Paramagul, et al., "Computer Aided Detection Systems for Breast Masses: Comparison of Performances on Full-Field Digital Mammograms and Digitized Screen-film Mammograms," Academic Radiology 6, 659-669 (2007).

[22]J. Wei, B. Sahiner, H. Chan, M. A. Roubidoux, M. A. Helvie, Y. Wu, L. M. Hadjiiski, J. Ge, and C. Zhou, "Computer-aided detection of breast masses on mammograms: Performance improvement using two-view information," Presentation at the 92nd Scientific Assembly and Annual Meeting of the Radiological Society of North America, Chicago, IL. November 26-December 1, 2006.

[23]D. P. Chakraborty and L. H. L. Winter, "Free-response methodology: Alternate analysis and a new observer-performance experiment," Radiology 174, 873-881 (1990).

[24]D. P. Chakraborty and K. S. Berbaum, "Observer studies involving detection and localization: modeling, analysis, and validation," Medical Physics 31, 2313-2330 (2004).

(10) Appendix

Copies of the following publications are enclosed with this report:

Journal Articles:

1. Wei J, Chan HP, Sahiner B, Hadjiiski LM, Helvie MA, Roubidoux MA, Zhou C, Ge J, "Dual system approach to computer-aided detection of breast masses on mammograms", Medical Physics, Vol. 33, No. 11, pp. 4157–4168, 2006.

2. Wei J, Hadjiiski LM, Sahiner B, Chan HP, Ge J, Roubidoux MA, Helvie MA, Zhou C, Wu YT, Paramagul C, Zhang Y, “Computer Aided Detection Systems for Breast Masses: Comparison of Performances on Full-Field Digital Mammograms and Digitized Screen-film Mammograms”, Academic Radiology, Vol. 14, No. 6, pp. 659-669, 2007.

3. Wu YT, Wei J, Hadjiiski LM, Sahiner B, Zhou C, Ge J, Shi J, Zhang Y, Chan HP, “Bilateral analysis based false positive reduction for computer-aided mass detection”, Medical Physics (in press).

Conference Proceeding:

1. Wei J, Sahiner B, Chan HP, Hadjiiski LM, Roubidoux MA, Helvie MA, Ge J, Zhou C, and Wu YT, “Computer-aided detection of breast masses on prior mammograms”, Proc. SPIE, Vol. 6514, pp. 51-57, 2007.

Dual system approach to computer-aided detection of breast masses on mammograms

Jun Wei,[a] Heang-Ping Chan, Berkman Sahiner, Lubomir M. Hadjiiski, Mark A. Helvie, Marilyn A. Roubidoux, Chuan Zhou, and Jun Ge
Department of Radiology, University of Michigan, Ann Arbor, Michigan 48109

(Received 13 November 2005; revised 30 August 2006; accepted for publication 31 August 2006; published 18 October 2006)

In this study, our purpose was to improve the performance of our mass detection system by using a new dual system approach which combines a computer-added detection (CAD) system optimized with "average" masses with another CAD system optimized with "subtle" masses. The two single CAD systems have similar image processing steps, which include prescreening, object segmentation, morphological and texture feature extraction, and false positive (FP) reduction by rule-based and linear discriminant analysis (LDA) classifiers. A feed-forward backpropagation artificial neural network was trained to merge the scores from the LDA classifiers in the two single CAD systems and differentiate true masses from normal tissue. For an unknown test mammogram, the two single CAD systems are applied to the image in parallel to detect suspicious objects. A total of three data sets were used for training and testing the systems. The first data set of 230 current mammograms, referred to as the average mass set, was collected from 115 patients. We also collected 264 mammograms, referred to as the subtle mass set, which were one to two years prior to the current exam from these patients. Both the average and the subtle mass sets were partitioned into two independent data sets in a cross validation training and testing scheme. A third data set containing 65 cases with 260 normal mammograms was used to estimate the FP marker rates during testing. When the single CAD system trained on the average mass set was applied to the test set with average masses, the FP marker rates were 2.2, 1.8, and 1.5 per image at the case-based sensitivities of 90%, 85%, and 80%, respectively. With the dual CAD system, the FP marker rates were reduced to 1.2, 0.9, and 0.7 per image, respectively, at the same case-based sensitivities. Statistically significant ($p < 0.05$) improvements on the free response receiver operating characteristic curves were observed when the dual system and the single system were compared using the test sets with either average masses or subtle masses. [DOI: 10.1118/1.2357838]

Key words: computer-aided detection (CAD), mass detection, mammogram, dual system, artificial neural network (ANN)

I. INTRODUCTION

Breast cancer is one of the leading causes of cancer mortality among women.[1] It has been reported that early diagnosis and treatment can significantly improve the chance of survival for patients with breast cancer.[2–4] At present, the most successful method for the early detection of breast cancer is screening mammography.[5] Various methods are being developed to improve the accuracy of breast cancer detection. Double reading by radiologists can reduce the miss rate of radiographic reading. However, double reading will increase the cost of mammographic screening. An alternative method is to use a trained computer-aided detection (CAD) system as a second reader.[6,7] Recent clinical studies have shown that CAD systems are helpful for increasing radiologists' accuracy in detecting breast cancers.[8–13]

A large volume of literature has been published in the CAD area. CAD systems for mammography generally consist of two subsystems: one is a mass detection system and the other is a microcalcification detection system. Detection of masses on mammograms is often more challenging than detection of microcalcifications. The mass detection systems to-date have employed a single-system approach using various techniques for prescreening of mass candidates and classification of true and false positives.[14–24] Our laboratory incorporated two-view mammographic information for improved differentiation of true masses and false positives and obtained promising preliminary results.[22] However, development of new methods to improve the performance of mass detection systems remains an important area of CAD research.

The CAD systems developed so far have mostly used masses seen on current mammograms (i.e., the mammograms on which the masses were detected by radiologists) for training. An important purpose of a CAD system is that it is used as a second reader to alert radiologists to subtle cancers that may be overlooked. To study the ability of a CAD system in detecting subtle cancers that are likely to be missed by radiologists, one way is to evaluate its accuracy in detecting missed cancers on prior mammograms (i.e., the mammograms in previous examinations on which the mass or cancer can be seen retrospectively but was considered

negative or benign at the time of the examination). Some researchers have investigated the performance change of CAD systems when using prior mammograms as input. In our study of mass detection on prior mammograms,[25] we obtained a case-based sensitivity of 74% (20/27) of the malignant masses with 2.2 false positives (FPs) per image. te Brake *et al.*[26] reported that their CAD system has a case-based sensitivity of 34% (22/65) of the cancers which have the appearance of masses or stellate lesions in the prior examinations with 1 FP per image. A commercial system (R2 ImageChecker) also reported detection of 42% (72/172) of the cancers in the prior years which were considered worthy of call-back in retrospect by expert mammographers with about 2 FP marks/case.[27] Zheng *et al.*[23] reported that their CAD system trained with current mammograms could not perform optimally in prior mammograms and vice versa; whereas the same system trained with prior mammograms can perform better on detecting the masses on prior mammograms. Recently, an assessment study[28] was conducted to compare the performance of two commercial systems and one research CAD system on current mammograms and prior mammograms. The results showed that the true positive (TP) fraction for CAD systems on prior mammograms of 39 breasts with malignant masses ranged from 15% to 26% with 0.28 to 0.41 FP marks/image. Although the detection performance reported in the different studies vary, probably due to the differences in the data set used, these studies indicate that the sensitivities of current CAD systems in detecting subtle masses on prior mammograms are substantially lower than that obtained from detection on current mammograms. The difficulty in recognizing the subtle and possibly different features of the masses on priors compared to those of the masses on current mammograms may be one of the factors that causes oversight for both radiologists and the CAD systems.

The goal of pattern recognition is to achieve the best possible classification performance in the task at hand. Researchers had shown that, for a class of objects with a wide range of characteristics, the classification performance can be improved by using combination of classifiers whereby objects of certain characteristics are classified by one classifier using a set of features and objects of different characteristics by another classification scheme based on different features.[29–35] The advantage of using combination of classifiers is that it may stabilize the training of classifiers even with a relatively small sample size because each classifier does not have to accommodate a wide range of characteristics and features.[36,37] These observations motivated our interest in the design of a dual CAD system for mass detection.

Since the missed cancers on prior mammograms represent the difficult cases that are more likely to be missed by radiologists if similar cancers occur on screening mammograms, it is important to improve the sensitivity of the CAD system in detecting these cancers. On the other hand, when a CAD system is applied to a new mammogram in clinical practice, it has to detect breast lesions of all degrees of subtlety effectively. However, it is difficult to train a single CAD system to provide optimal detection for all lesions over the entire spectrum of subtlety because the classifiers have to make compromises to accommodate cancers of a wide range of characteristics. Therefore, we have been exploring a new dual CAD system approach that combines a CAD system trained with retrospectively seen masses on prior mammograms with a CAD system trained with masses detected on current mammograms.[38,39] In this paper, we will describe the design of the dual CAD system and report our current results.

II. MATERIALS AND METHOD

A. Data sets

All mammograms in this study were collected from patient files in the Department of Radiology at the University of Michigan with Institutional Review Board (IRB) approval. The mammograms were digitized with a LUMISYS 85 laser film scanner with a pixel size of 50 μm $\times$ 50 μm and 4096 gray levels. The scanner was calibrated to have a linear relationship between gray levels and optical densities (O.D.) from 0.1 to greater than 3 O.D. units. The nominal O.D. range of the scanner is 0–4. The full resolution mammograms were first smoothed with a 2×2 box filter and subsampled by a factor of 2, resulting in 100 μm $\times$ 100 μm images. The images at a pixel size of 100 μm $\times$ 100 μm were used for the input of our CAD system.

We collected three data sets. The first data set contained 115 cases with confirmed masses. Each case included the current mammograms that prompted the radiologist to work up the mass. This is referred to as the "average" mass set. All of the cases in the average mass set had two mammographic views: the craniocaudal view and the mediolateral oblique view or the lateral view, thus yielding a total of 230 mammograms. There were 115 masses (67 malignant masses and 48 benign masses) in this data set, of which 105 were biopsy-proven and 10 were determined to be benign by long-term follow-up.

The second data set was composed of the prior mammograms dated one to two years earlier than the mammograms of the same patients in the average mass set. Since the masses on prior mammograms are on average subtler than those on current mammograms, this data set is referred to as the "subtle" mass set. On 5 of the 115 patients, no mass or focal density could be identified on either view of the prior mammograms. Therefore, the subtle mass set was composed of 110 cases (62 malignant and 48 benign). For the purpose of training the subtle mass detection system, the subtle masses do not have to be obtained from the same cases as the average mass set but we used the available prior mammograms for these mass cases in our database. Nineteen of the 110 cases had two prior mammogram examinations. Of the 129 examinations in the subtle mass set, 123 had two mammographic views and 6 had three views, with a total of 264 mammograms. Many of the subtle masses on the prior mammograms could be identified only as a focal density corresponding to the location of the subsequently detected mass on the current mammograms. On 44 of the two-view prior

TABLE I. Description of cases in the average and subtle mass data sets and the subsets for training and testing in the cross-validation scheme.

	Mass subset 1		Mass subset 2	
	Average mass subset	Subtle mass subset	Average mass subset	Subtle mass subset
Total No. of cases	57	54	58	56
Cases with two prior examinations	NA	10	NA	9
Exams with two views	57	58	58	65
Exams with three views	0	6	0	0
Total No. of images	114	134	116	130
No. of negative images	0	25	0	19
No. of mass images for training	114	109	116	111
No. of two-view pairs for testing	57	64	58	65
No. of images for testing	114	128	116	130
No. of malignant masses	36	33	31	29
No. of benign masses	21	21	27	27

mammograms, the mass location was evident only on one view. Table I summarizes the information for the average and subtle mass subsets.

The third data set was composed of 260 normal bilateral two-view mammograms obtained from 65 patients. No masses were evident on these mammograms upon review by the experienced radiologist.

The two mass data sets were used to estimate the detection sensitivity and the normal data set was used for estimating the FP marker rate. For the mass data sets, the true locations of the masses were identified by an experienced MQSA radiologist using all available imaging and clinical information. The radiologist also provided an estimate of the longest diameter of the mass, descriptors of its margin and shape, a visibility rating, and an estimate of the breast density in terms of BI-RADS category. Figure 1 shows the distributions of mass sizes, mass shapes, mass margins, and their visibility on a 10-point rating scale with 1 representing the most visible masses and 10 the most difficult case relative to the cases seen in their clinical practice. The masses had a mean of 13.7 mm and a median of 12 mm in the average data set and a mean of 9.7 mm and a median of 10 mm in the subtle data set. Figure 2 shows the breast density for both the normal data set and the mass data sets. As can be seen from the distributions of the mass characteristics, the average masses on the current mammograms and the subtle masses on the priors had large overlap. Nevertheless, on average, the subtle masses were smaller in size and less conspicuous on the mammograms.

B. Methods

In order to improve the sensitivity of detecting breast lesions of all degrees of subtlety, we developed a new dual system approach which combines a system trained with average masses with another system trained with subtle masses. When the trained dual system is applied to an unknown mammogram, the two CAD systems are used in parallel to detect suspicious objects on a single mammogram. No prior mammogram is needed. The additional FPs from the use of the two systems are reduced by an information fusion stage. We will refer to the two systems separately trained with the average masses and the subtle masses as "single" CAD systems in the following discussions.

We randomly separated the mass data sets by case into two independent subsets. Both the average and subtle mass subsets followed the same case grouping so that mammograms from the same case would not be separated into the training subset for one single CAD system and the test subset for the other single CAD system in a cross-validation cycle. Table I shows the subsets of cases in the average and subtle mass data sets. Two-fold cross validation was used for training and testing the algorithms. The training included selecting proper parameters for each single CAD system and for information fusion. Once the training with one mass subset was completed, the parameters were fixed for testing with the other mass subset. The training and test mass subsets were switched and the training and test processes were repeated. The CAD systems were trained with single mammograms. To maximize the number of training images with masses, all images with a visible mass were included regardless of whether they were a part of a two-view or three-view case when the subtle mass subset was used as a training set. However, when the subtle mass subset was used as a test set, only two views were included for each case because we used two-view mammograms to derive the case-based test performance. For cases containing three views, we therefore included only two of the views in testing. We also included cases with the mass visible on only one of the two views. After the two-fold cross validation testing, the overall detection performance was evaluated by combining the performances of the two test subsets. The trained algorithms with the fixed parameters were also applied to the normal set of mammograms, which was not used during training, to estimate the FP rate in screening mammograms.

1. Single CAD system overview

The major steps in the two single mass detection systems are similar but the feature spaces and classifiers for FP reduction in each system were designed separately to suit the characteristics of average and subtle masses, respectively. The two systems are therefore described together in the following but the differences will be pointed out whenever applicable. Each single CAD system consists of four processing steps: (1) prescreening of mass candidates, (2) segmentation of suspicious objects, (3) feature extraction and

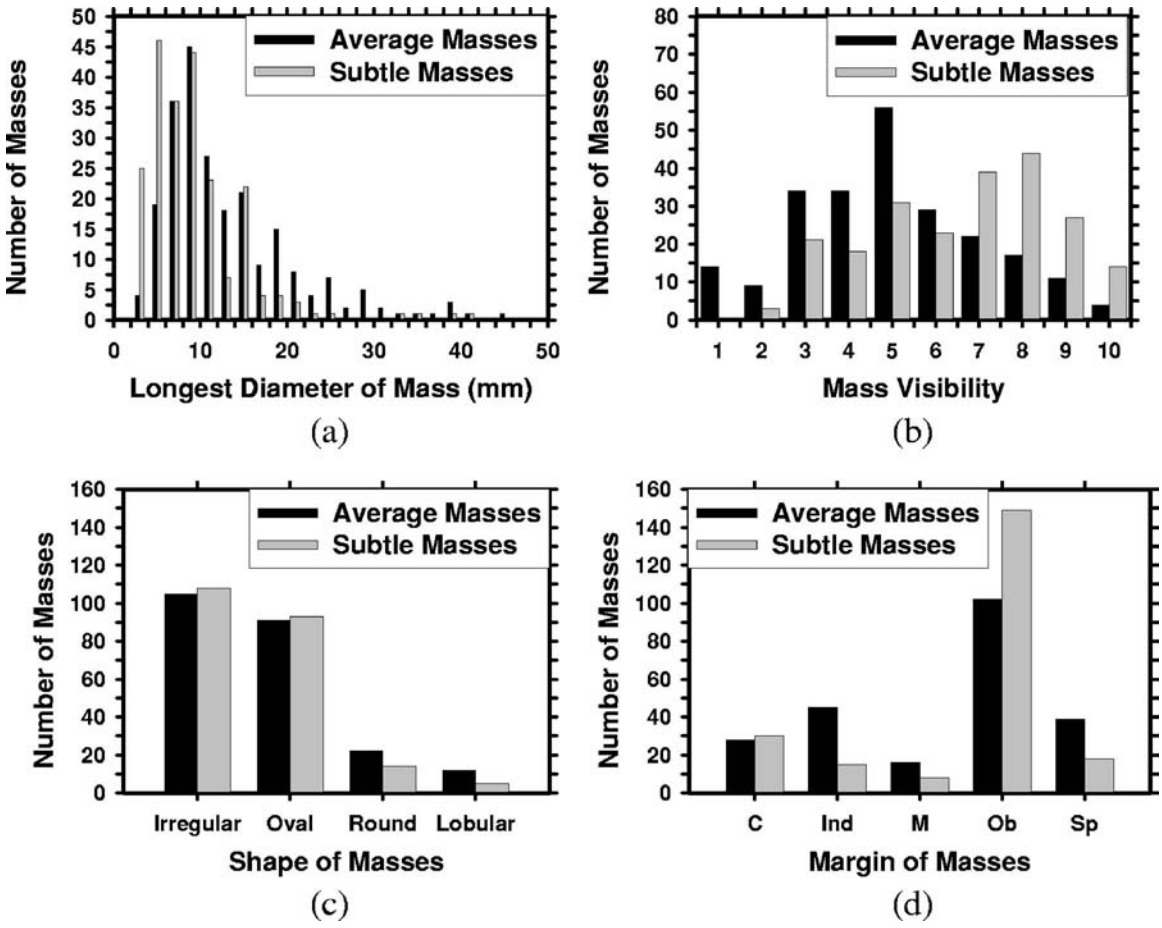


FIG. 1. The characteristics of the masses in our mass data set: (a) distribution of mass sizes, (b) distribution of mass visibility on a 10-point rating scale with 1 representing the most visible masses and 10 the most subtle masses relative to the cases seen in clinical practice, (c) distribution of mass shapes, (d) distribution of mass margins, C: circumscribed, Ind: indistinct, M: microlobulated, Ob: obscured, Sp: spiculated.

analysis, and (4) FP reduction by classification of normal tissue structures and masses. The block diagram for the detection scheme is shown in Fig. 3.

For the prescreening stage, we have developed a two-stage gradient field analysis method which not only uses the shape information of masses on mammograms but also incorporates the gray level information of the local object segmented by a region growing technique in the second stage to refine the gradient field analysis.[24,40] Locations of high radial gradient convergence are labeled as mass candidates. After prescreening, the suspicious objects are identified by using a two-stage segmentation method.[41] First, the background-

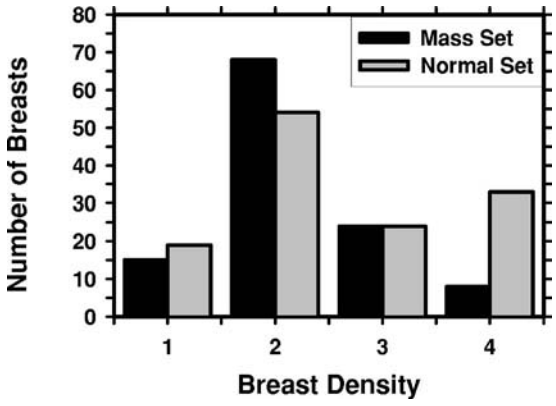


FIG. 2. The distribution of breast density in terms of BI-RADS categories estimated by an MQSA radiologist.

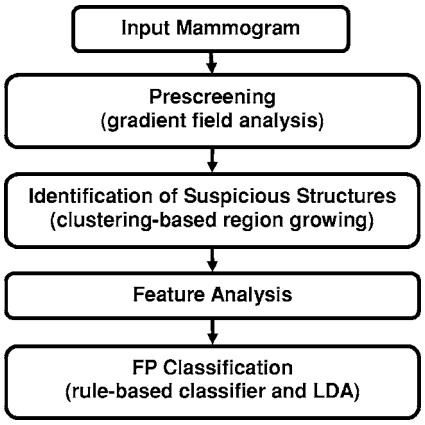


FIG. 3. Schematic diagram of our single CAD system for mass detection. The FP classification stage includes rule-based classification, a morphological LDA classifier, and a texture feature LDA classifier for differentiating masses from normal breast tissues.

corrected ROI is weighted by a two-dimensional Gaussian function with $\sigma=256$ pixels to enhance the central region. Sobel filtering is then applied to the Gaussian-weighted ROI to generate another enhanced image. Second, a *k*-means clustering using the pixel values from these two images as features is used to segment the object. For each suspicious object, eleven morphological features[21] were extracted. Rule-based and linear discriminant classifiers were trained by using the training data set only to remove the detected structures that were substantially different from breast masses. For the system trained with average masses, global and local multiresolution texture analysis[42] were performed in each ROI by using the spatial gray level dependence (SGLD) matrices. A total of 364 features were extracted from global texture analysis. Local texture features were extracted from the local region containing the detected object and the peripheral regions within each ROI. A total of 208 features were extracted for local texture analysis. For the system trained with subtle masses, instead of the SGLD texture features, gray level features and run length statistics analysis (RLS) texture features[43] were extracted inside and outside of each mass region on the original image and gradient field image. The gray level features included the contrast of the object relative to the surrounding background, the minimum and the maximum gray levels, and the characteristics derived from the gray level histogram in the regions inside and outside of each object including skewness, kurtosis, energy, and entropy. Five RLS texture features were extracted in both the horizontal and vertical directions: short runs emphasis, long runs emphasis, gray level nonuniformity, run length nonuniformity, and run percentage. A total of 66 features were extracted for the system trained with subtle masses.

In order to obtain the best texture feature subset and also reduce the dimensionality of the feature space to design an effective classifier, stepwise feature selection with linear discriminant analysis (LDA) was applied to the training subset. The detailed procedure has been described elsewhere.[24,44,45] Briefly, at each step one feature was entered or removed from the feature pool by analyzing its effect on the selection criterion, which was chosen to be the Wilks' lambda in this study. Since the appropriate values of thresholds for feature entry, feature elimination, and tolerance of correlation for feature selection were unknown, we used an automated simplex optimization method to search for the best combination of thresholds in the parameter space. The simplex algorithm used a leave-one-case-out resampling method within the training subset to select features and estimate the weights for the LDA classifier. To have a figure-of-merit to guide feature selection, the test discriminant scores from the left-out cases were analyzed using receiver operating characteristic (ROC) methodology.[46] The accuracy for classification of masses and FPs was evaluated as the area under the ROC curve, A_z. In this approach, feature selection was performed without the left-out case so that the test performance would be less optimistically biased.[47] However, the selected feature set in each leave-one-case-out cycle could be slightly different because every cycle had one training case different from the other cycles. In order to obtain a single trained classifier to apply to the independent test subset, a final stepwise feature selection was performed with the best combination of thresholds, found in the simplex optimization procedure, on the entire training subset to obtain the final set of features and estimate the weights of the LDA. Note that the entire process of feature selection and classifier weight estimation was performed within the training subset. The LDA classifier with the selected feature set was then fixed and applied to the independent test subset. The training and testing processes were performed independently for the two-fold cross-validation sets.

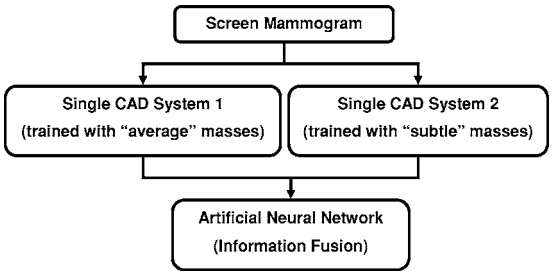


FIG. 4. Schematic diagram of proposed dual CAD system for mass detection. BP-ANN is used for information fusion.

2. Training and test for dual system

The block diagram for the dual system is shown in Fig. 4. During the training of the dual system, we used the current and prior mammograms from the same patients. The current mammograms that contained the average masses were only used to train the first single CAD system. The prior mammograms that contained the subtle masses were only used to train the second single CAD system. The prescreening and the segmentation steps in the two systems are identical. Since the morphological appearances of average and subtle masses are different, the rules in the morphological rule-based FP classification are trained differently for the two single CAD systems. During testing with an independent mammogram, the dual system keeps all the suspicious objects that satisfy the FP classification rules of either single CAD system and applies the LDA classifiers from both single systems to each object. Each object thus has two LDA scores.

To merge the information from the two CAD systems, a fusion scheme was developed for our dual system. In this study, a feed-forward backpropagation artificial neural network (BP-ANN) was trained to classify the masses from normal tissues by combining the output information from the two single CAD systems. The LDA classifiers from the two single CAD systems were applied to each detected object. The two LDA discriminant scores for each object were used as input to the BP-ANN. The BP-ANN had an input layer with two nodes, a hidden layer with N nodes, and an output layer with one node. The nodes were interconnected by weights and information propagated from one layer to the next through a log-sigmoidal activation function. The learning of the ANN was a supervised process in which known training cases were input to the ANN. The performance func-

tion for the network was the mean-squared error between the network outputs and the target outputs. The weights of the network were adjusted iteratively by a feedforward backpropagation procedure to minimize the error. Detailed description of the backpropagation neural network can be found in the literature.[48,49]

To choose the number of hidden nodes (N) in the BP-ANN, we used a three-fold cross-validation method within the training subset. We randomly separated the entire training subset including all detected objects into three independent groups. The objects belonging to the same case were separated into the same group. For a given N, three training cycles were performed, in each of which two of the three groups were used to train the BP-ANN and the left-out group was used to test its performance. The A_z value obtained from the ANN output scores for the test group was used as the performance index for that training cycle. The average of the A_z values from the three test groups represented the performance of the BP-ANN with N hidden nodes. In our experiment, a BP-ANN with 3 hidden nodes provided the largest average A_z value and was therefore chosen. The weights of the chosen BP-ANN were retrained with the entire training subset. The BP-ANN with the trained weights was used to merge the information from the two single CAD systems.

To test the dual system, the two trained single CAD systems, one trained with the average mass set and the other with the subtle mass set, were applied in parallel to each single "unknown" mammogram in the independent test subset. No prior mammogram was needed during testing.

3. Evaluation methods

The detected individual objects were compared with the "truth" ROI marked by the experienced radiologist, as described earlier. A detected object was scored as TP if the overlap between the bounding box of the detected object and the bounding box of the true mass relative to the larger of the two bounding boxes was over 25%. Otherwise, it would be scored as FP. The 25% threshold was selected as described in our previous study.[21]

The FP marker rate was estimated in two ways: one from detection on the same test subsets with masses, the other from detection on the normal data set of negative mammograms. For the latter, we applied the trained dual CAD system to the normal data set. The number of FP marks produced by the CAD system was determined by counting the detected objects on the normal cases. The mass detection sensitivity was determined by counting the detected masses on the test mass subset. The detection performance of the CAD system was assessed by free response ROC (FROC) analysis. A FROC curve was obtained by plotting the mass detection sensitivity as a function of FP marks per image either obtained from the mass data subset or the normal set at the corresponding decision threshold.

FROC curves were presented on a per-mammogram and a per-case basis. For image-based FROC analysis, the mass on each mammogram was considered an independent true object. For case-based FROC analysis, the same mass imaged on the two-view mammograms was considered to be one true object and detection of either or both masses on the two views was considered to be a TP detection.

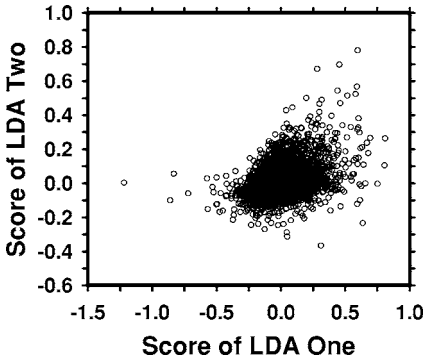


FIG. 5. An example of a scatter plot of the LDA scores from the two single CAD systems which are used as input to the BP-ANN. The correlation coefficient between the scores of two LDA classifiers is 0.46, indicating that the two LDA scores are essentially independent features.

Since we used two-fold cross validation method for training and testing, we obtained two test FROC curves, one for each test subset, for each of the conditions (e.g., single CAD system approach or dual system approach). To summarize the results for comparison, an average test FROC curve was derived by averaging the FP rates at the same sensitivity along the FROC curves of the two corresponding test subsets.

In order to compare the performance of the single CAD system and the dual CAD system, we applied the alternative free-response ROC (AFROC) method and the jackknife free-response ROC (JAFROC) method developed by Chakraborty *et al.*[50,51] to the pairs of FROC curves. In the AFROC method, the FROC data are first transformed by counting the number of false-positive images (FPIs) instead of the FPs per image. The confidence rating of a FPI is determined by the highest confidence FP decision on the image regardless of how many lower confidence FP decisions are made on the same image. The ROCKIT curve fitting software and statistical significance tests for ROC analysis developed by Metz *et al.*[46] can then be used to analyze the AFROC data.

III. RESULTS

Figure 5 shows an example of the two-dimensional feature space that was used as the input to the BP-ANN being trained to merge the information from the two single CAD subsystems. The two features are the output scores of the LDA classifiers trained with the average masses and with the subtle masses. The correlation coefficients of the two features are 0.46 and 0.44 for each of the training subsets, respectively. The low correlation indicated that the two single CAD systems extracted relatively independent features from the object. The A_z values of the chosen ANN were 0.92 ± 0.01 and 0.87 ± 0.01, respectively, as estimated by validation in the training process. The ANN classifiers achieved A_z values

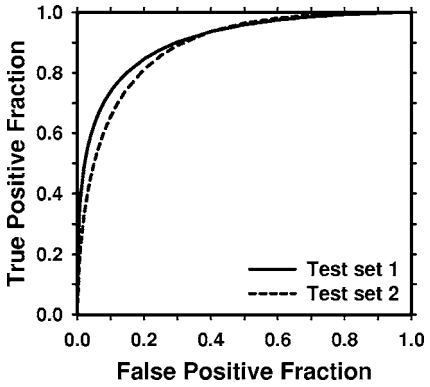


FIG. 6. The test ROC curves for the BP-ANN classifiers from the two independent mass subsets. The ANN classifiers achieved an A_z value of 0.90±0.02 for test subset 1 and 0.89±0.01 for test subset 2 in the classification of mass and normal breast tissues.

of 0.90±0.02 and 0.89±0.01 on the two independent test subsets, respectively. Figure 6 shows the ROC curves for the two test subsets.

In order to evaluate the effectiveness of our dual system approach, we compared its performance on the test subsets containing average masses with two other single CAD systems: the CAD system trained only on the average mass set and the CAD system trained on both the average and the subtle mass sets. When a single CAD system was trained only with the average masses, the number of selected features was 21 (14 global and 7 local) and 16 (10 global and 6 local) texture features for the two independent training subsets, respectively. When the CAD system was trained with both the average and the subtle masses, the number of selected features was 17 (11 global and 6 local) and 18 (7 global and 11 local) texture features for the two independent training subsets, respectively.

For the dual system, the single system trained with the average masses was the same as that described earlier. For the single system trained with subtle masses, four (2 gray level and 2 RLS texture) and five (3 gray level and 2 RLS texture) features were selected for the two independent training subsets, respectively.

The average test FROC curves of the dual CAD system on the test subsets with average masses were compared to those of the single CAD systems in Fig. 7. The FP rates were estimated from the mass data set. The dual CAD system achieved a case-based sensitivity of 80%, 85%, and 90% at 0.6, 0.8, and 1.0 FPs/image, respectively, compared with 1.3, 1.5, and 1.8 FPs/image on the single CAD system trained with average masses alone. The performance of the single CAD system trained with both the average masses and the subtle masses was comparable to that trained with average masses alone, with FP rates of 1.4, 1.6, and 1.8 FPs/image at the same sensitivities, respectively. Figure 8 shows the comparison of the three average test FROC curves, similar to those shown in Fig. 7, except that the FP rates were estimated from the normal data set. The FP rates at a few selected sensitivities for the dual and single CAD systems were summarized in Table II.

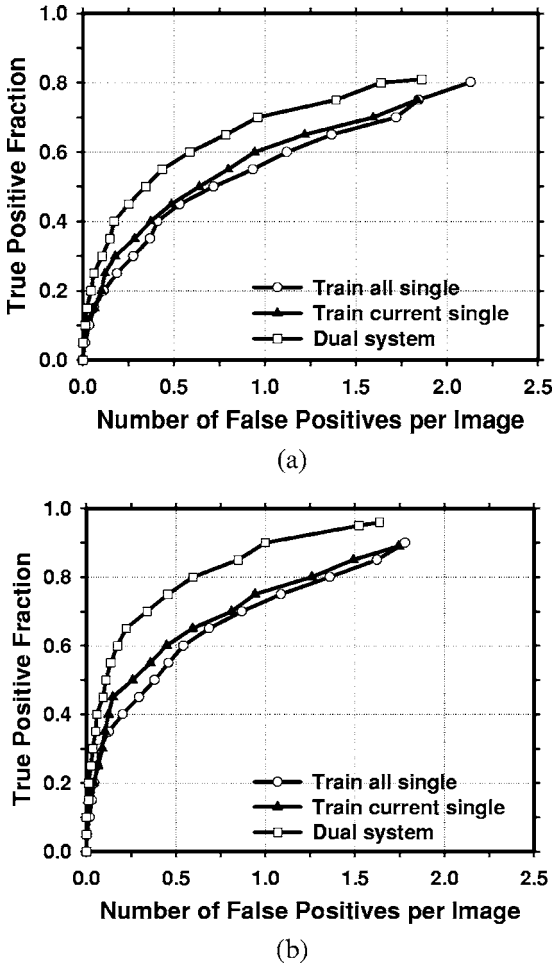


FIG. 7. Comparison of the average test FROC curves obtained from averaging the FROC curves of the two independent average-mass subsets. Three CAD systems were compared: a single CAD system trained with average masses alone, a single CAD system trained with both the average and the subtle masses, and the dual CAD system. The FP rate was estimated from the mammograms with masses. (a) Image-based FROC curves, (b) case-based FROC curves.

In this study, we have 67 malignant cases in the average mass set. Figure 9 compares the average test FROC curves of the single CAD system and the dual system for detection of malignant masses. The result for the single CAD system trained with average masses was shown and the FP rate was estimated from the mammograms without masses. In this case, the dual CAD system achieved a case-based sensitivity of 80%, 85%, and 90% at 0.6, 0.9, and 1.2 FP marks/image, respectively, compared with 1.1, 1.6, and 2.0 FP marks/image on the single CAD system.

An important purpose of a CAD system is to serve as a second reader to alert radiologists to subtle cancers that may be overlooked. Figures 10 and 11 compare the average FROC curves of the single CAD system and the dual system for detection in the test subsets with subtle masses. The TP rate in Fig. 10 was estimated by including both malignant and benign masses and that in Fig. 11 was estimated from malignant masses only. The single CAD system trained with average masses alone was used. The FP rates for both sys-

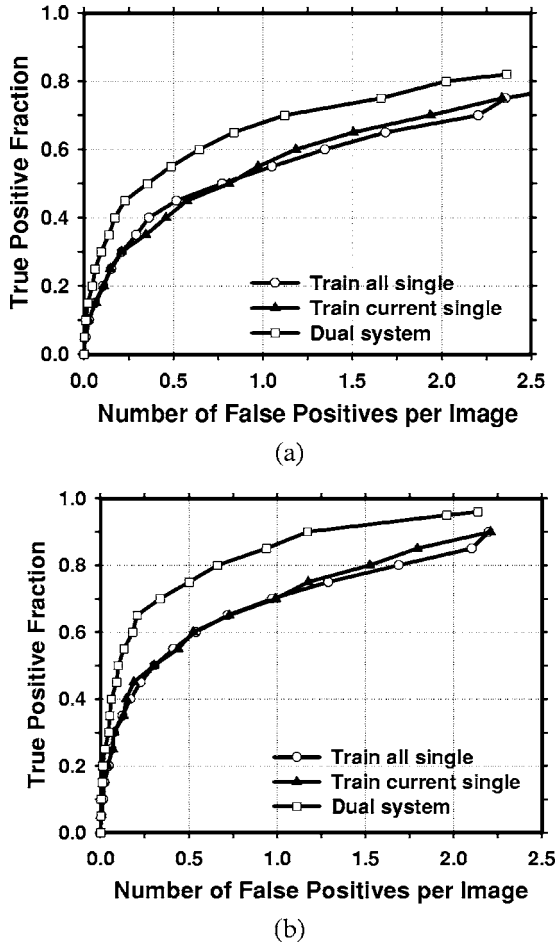


FIG. 8. Comparison of the average test FROC curves obtained from averaging the FROC curves of the two independent average-mass subsets. Three CAD systems were compared: a single CAD system trained with average masses only, a single CAD system trained with the average and the subtle masses, and the dual CAD system. The FP rate was estimated from the mammograms without masses. (a) Image-based FROC curves, (b) case-based FROC curves.

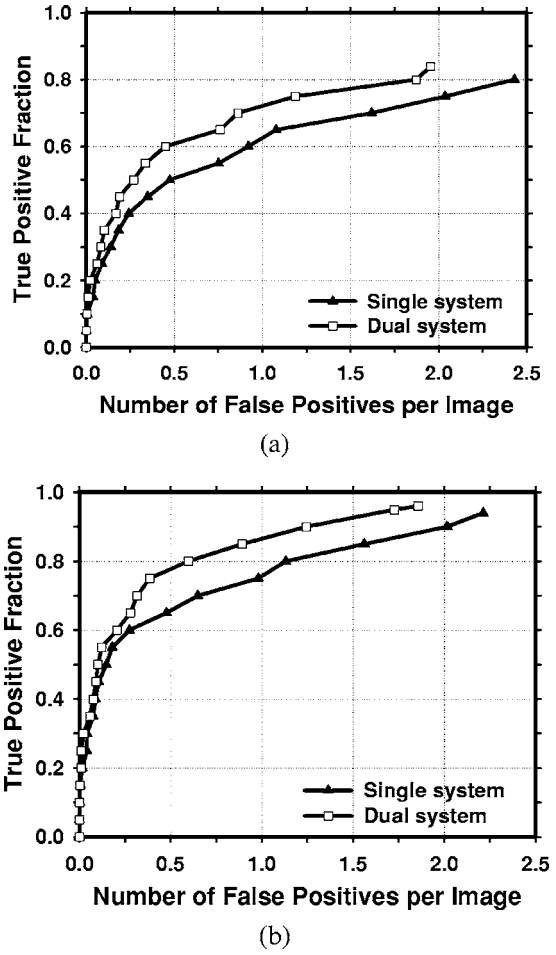


FIG. 9. Comparison of the average test FROC curves of the single CAD system and the dual CAD system for detection of malignant masses in the average data set. The single system trained with average masses alone was used and the FP rate was estimated from the mammograms without masses. (a) Image-based FROC curves, (b) case-based FROC curves.

tems were estimated from the mammograms without masses. The dual CAD system achieved a case-based sensitivity of 50% at 0.7 FP marks/image for all masses and at 0.5 FP marks/image for malignant masses only, compared with 1.4 FP marks/image for all masses and 1.1 FP marks/image for malignant masses only using the single CAD system.

TABLE II. Comparison of case-based detection performance between the dual system and the single CAD system trained with average masses alone. The FP marker rates were estimated from detection on the normal data set. The FROC curves were obtained by averaging the FROC curves of the two test subsets.

	Average mass test set (FP marks/image)		Subtle mass test set (FP marks/image)	
TP	Single system	Dual system	Single system	Dual system
90%	2.2	1.2		
80%	1.5	0.7		2.8
70%	1.0	0.3	2.4	2.3
60%	0.5	0.2	1.8	1.5
50%	0.3	0.1	1.4	0.7

Table II summarizes the test results on the average and subtle mass sets for the dual system and the single CAD system trained with average masses at different sensitivity levels. The FP marker rates were estimated from the detection on the normal data set.

The comparison of the FROC curves for the dual CAD system and the single CAD system in terms of the area under the fitted AFROC curve (A_1) and the p values for both test subsets with average masses was summarized in Table III. The differences between the A_1 values for the two systems were statistically significant ($p<0.05$). The fitted AFROC curves, however, did not fit very well to the transformed AFROC data, as we discussed previously.[24] For the JAFROC method, Chakraborty *et al.* provided software to estimate the statistical significance of the difference between two FROC curves. The comparison of the figure-of-merit (FOM) and the p values was also summarized in Table III. The differences

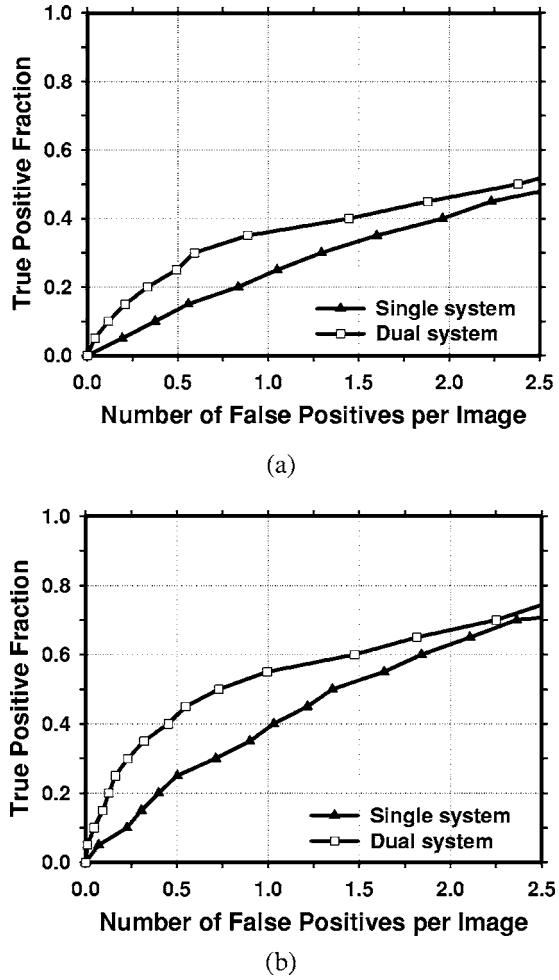


FIG. 10. Comparison of the average test FROC curves for the single CAD system and the dual CAD system for detection of the subtle masses on the prior mammograms. The single CAD system trained with average masses alone was used and the FP rate was estimated from the mammograms without masses. (a) Image-based FROC curves, (b) case-based FROC curves.

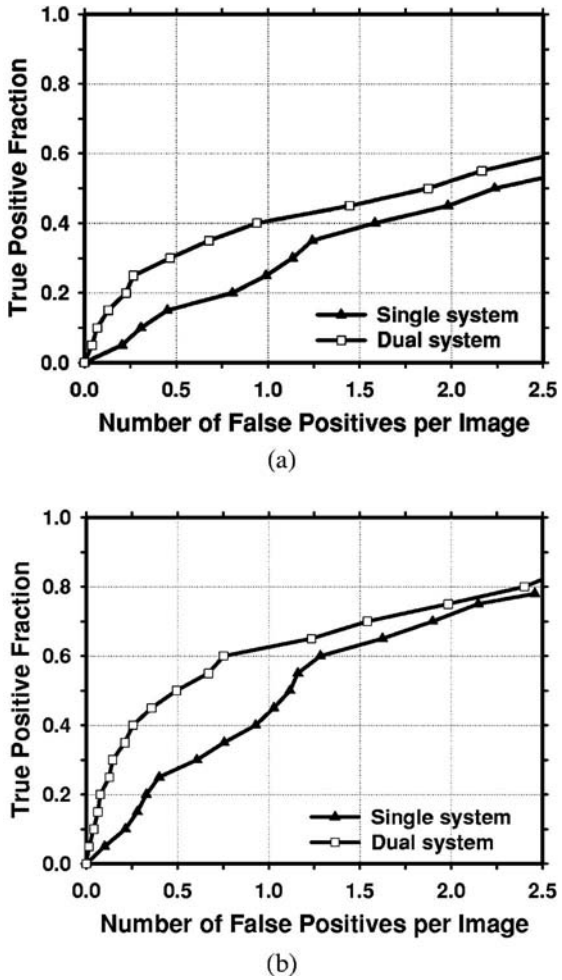


FIG. 11. Comparison of the average test FROC curves for the single CAD system and the dual CAD system for detection of subtle malignant masses on the prior mammograms. The single CAD system trained with average masses alone was used and the FP rate was estimated from the mammograms without masses. (a) Image-based FROC curves, (b) case-based FROC curves.

between the FOM of the dual CAD system and that of the single CAD system for both test subsets were again statistically significant ($p < 0.05$).

The comparison of A_1, the FOM, and the p values for the dual system and the single system trained with average masses in detecting subtle masses was summarized in Table IV. It was found that the differences between the results of the dual CAD system and those of the single CAD system on the two test subsets containing subtle masses were statistically significant by both the JAFROC and the AFROC methods.

IV. DISCUSSION

The masses on prior mammograms are more subtle and more difficult to detect than the masses on current mammograms. In this study, we developed a dual CAD system, which combines a system trained with masses on prior mammograms and a system trained with masses detected on current mammograms. We have demonstrated that this dual system can increase the accuracy of detecting both average masses and subtle masses. The comparisons of the dual system with that of the single CAD system trained with average masses alone and that of the single CAD system trained with both average and subtle masses (Fig. 7) indicate that the gain in the detection accuracy of the dual system could not be achieved by simply using a larger training set with both average and subtle masses. In fact, it is interesting to note that the performance of the single CAD system trained with both the average and the subtle masses appeared to be degraded slightly, in comparison with the single system trained with average masses alone, when it was applied to the test set of average masses. The decreased performance may reflect the compromise made when the single CAD system was trained to accommodate a wide range of lesion characteristics. Thus, the dual system approach may have improved its performance through other factors, including the flexibility in using different feature spaces and training the parameters for each type of masses and the information fusion combining the two single CAD systems effectively.

TABLE III. Estimation of the statistical significance in the difference between the FROC performance of the dual system and the single CAD system trained with average masses alone when the systems were evaluated on the average mass test subsets. The FROC curves with the FP marker rates obtained from the normal data set were compared.

	A_1 (AFROC)				FOM (JAFROC)			
	All cases		Malignant cases		All cases		Malignant cases	
	Test subset 1	Test subset 2	Test subset 1	Test subset 2	Test subset 1	Test subset 2	Test subset 1	Test subset 2
Single system	0.45	0.44	0.47	0.52	0.48	0.48	0.53	0.55
Dual system	0.55	0.53	0.58	0.62	0.60	0.56	0.63	0.64
p values	0.0004	0.0156	0.0003	0.0318	<0.0001	0.007	0.0004	0.0252

For the comparison of the different systems, we analyzed the false negatives (FNs) of the single CAD systems and the dual CAD system when the test subsets with average masses were used. It was found that the FN rates of the single CAD system trained with average masses, the single CAD system trained with subtle masses, and the dual system were 23.9% (55/230), 28.3% (65/230), and 16.5% (38/230), respectively, after FP reduction by the morphological LDA classifier in each system. Twenty-nine masses were missed by both of the single systems. By using the dual system, 53 masses that were FNs for either single system could be detected. However, the masses that were missed by both of the single CAD systems could not be recovered by the dual CAD system.

Our motivation of this study is to improve the performance of a CAD system for mass detection. A CAD detection system is generally intended for use in screening mammography. At the screening stage, all lesions of concern should be pointed out to radiologists so that the radiologists can judge if a recall is warranted. If a detection system is trained to mark only the malignant lesions, it may be attempting to play the role of a triage system (alerting radiologists to work up only "malignant" cases) rather than that of a second reader. Furthermore, since computerized lesion detection or characterization on mammograms is not 100% sensitive, it will be confusing to the radiologists whether an unmarked suspicious lesion is missed or it is considered benign by the computer. We believe that computer-aided diagnosis (CADx) may be used in different ways in conjunction with a CAD detection system, for example, the likelihood of malignancy may be estimated by the CADx system and displayed for every detected lesion, and/or a CADx system may be used during diagnostic workup. Either way the CAD system will first alert radiologists to all masses, leaving the assessment of malignancy or benignity to a second stage and with the radiologist being the primary decision maker. The training set thus included both malignant and benign masses.

For a CAD system, its performance for detecting malignant masses is more important than its performance for detecting all masses. The FROC curves for detection of malignant masses on the average data set and the subtle data set, shown in Figs. 9 and 11, respectively, indicated that the dual system could also achieve an improvement in the detection performance over that of the single system. The differences in the A_1 and the FOM for the detection of malignant cases in the average and subtle mass test subsets were statistically significant, as shown in Tables III and IV, respectively.

In screening mammography, the cancer rate is 3–5 per 1000. Most of the mammograms are normal. Therefore, some CAD researchers and users estimate the FP rate using

TABLE IV. Estimation of the statistical significance in the difference between the FROC performance of the dual system and the single CAD system trained with average masses alone when the systems were evaluated on the subtle mass test subsets. The FROC curves with the FP marker rates obtained from the normal data set were compared.

	A_1 (AFROC)				FOM (JAFROC)			
	All cases		Malignant cases		All cases		Malignant cases	
	Test subset 1	Test subset 2	Test subset 1	Test subset 2	Test subset 1	Test subset 2	Test subset 1	Test subset 2
Single system	0.17	0.20	0.24	0.25	0.21	0.23	0.24	0.26
Dual system	0.28	0.25	0.35	0.34	0.30	0.28	0.36	0.34
p values	<0.0001	0.046	<0.0001	0.0067	0.0007	0.048	<0.0001	0.0035

normal mammograms[52–54] because it reflects how the CAD system performs in terms of specificity and whether the CAD system may cause extra efforts for radiologists to double check the marked locations or unnecessary recalls in a screening setting. Furthermore, for CAD systems that set a maximum number of detected objects at the output, estimating the number of FPs using images with lesions can potentially lead to an optimistic bias for the FROC curve because one of the detected objects will likely be the true lesion. The FP rate can thus be underestimated by as much as 1 per image. In addition, the JAFROC analysis requires that the FP rates be estimated on normal images. We therefore reported the FP rates of our CAD systems on both mammograms with masses and without masses to facilitate comparison with other CAD systems in case investigators may evaluate their FP rates in either way.

In this study, we evaluated the performance of the trained CAD systems with an independent test set using the two-fold cross validation method. Although the selection of parameters and features was performed using the training set, we had full knowledge of the performance for the test set so that the selections could be optimistically biased. True independent testing will have to be performed with unknown cases that have never been used for testing the CAD system before, such as those in a prospective clinical trial. However, this test step is beyond the scope of our current developmental process. Since we used the same cross-validation method for evaluation of the dual system and the single CAD systems, the comparison of their relative performances is expected to be less biased than their individual performances.

V. CONCLUSION

We have proposed a new dual system approach which combines a system trained with subtle masses on prior mammograms and a system trained with average masses on current mammograms. The dual system achieved higher sensitivities at the corresponding FP rates than a single CAD system trained with average masses alone or trained with both average masses and subtle masses. Alternatively, the dual system had lower FP rates than the single CAD system at corresponding sensitivities. The improvement in the FROC curves by the dual system approach was found to be statistically significant ($p < 0.05$) for both average masses and subtle masses using either the AFROC or the JAFROC method. Our results indicate that the dual system approach is promising for improving the performance of CAD systems for mass detection on mammograms.

ACKNOWLEDGMENTS

This work is supported by U. S. Army Medical Research and Materiel Command Grant Nos. W81XWH-1-04-1-0475, DAMD 17-02-1-0214, and USPHS Grant No. CA95153. The content of this paper does not necessarily reflect the position of the government and no official endorsement of any equipment and product of any companies mentioned should be inferred. The authors are grateful to Charles E. Metz, Ph.D., for the LABROC program and to Dev Chakraborty, Ph.D., for the JAFROC program.

[a]Electronic mail: jvwei@umich.edu

[1]American Cancer Society, www.cancer.org, "Statistics for 2005."
[2]C. R. Smart, R. E. Hendrick, J. H. Rutledge, and R. A. Smith, "Benefit of mammography screening in women ages 40 to 49 years: Current evidence from randomized controlled trials," Cancer **75**, 1619–1626 (1995).
[3]S. A. Feig, C. J. D'Orsi, R. E. Hendrick, V. P. Jackson, D. B. Kopans, B. Monsees, E. A. Sickles, C. B. Stelling, M. Zinninger, and P. Wilcox-Buchalla, "American College of Radiology guidelines for breast cancer screening," AJR, Am. J. Roentgenol. **171**, 29–33 (1998).
[4]B. Cady and J. S. Michaelson, "The life-sparing potential of mammographic screening," Cancer **91**, 1699–1703 (2001).
[5]H. C. Zuckerman, "The role of mammography in the diagnosis of breast cancer," in *Breast Cancer, Diagnosis and Treatment*, edited by I. M. Ariel and J. B. Cleary (McGraw-Hill, New York, 1987).
[6]F. Shtern, C. Stelling, B. Goldberg, and R. Hawkins, "Novel technologies in breast imaging: National Cancer Institute perspective," Orlando, FL.
[7]C. J. Vyborny, "Can computers help radiologists read mammograms?," Radiology **191**, 315–317 (1994).
[8]T. W. Freer and M. J. Ulissey, "Screening mammography with computer-aided detection: Prospective study of 12,860 patients in a community breast center," Radiology **220**, 781–786 (2001).
[9]M. A. Helvie, L. M. Hadjiiski, E. Makariou, H. P. Chan, N. Petrick, B. Sahiner, S. C. B. Lo, M. Freedman, D. Adler, J. Bailey *et al.*, "Sensitivity of noncommercial computer-aided detection system for mammographic breast cancer detection—A pilot clinical trial," Radiology **231**, 208–214 (2004).
[10]R. L. Birdwell, P. Bandodkar, and D. M. Ikeda, "Computer-aided detection with screening mammography in a university hospital setting," Radiology **236**, 451–457 (2005).
[11]D. Gur, J. H. Sumkin, H. E. Rockette, M. A. Ganott, C. Hakim, L. A. Hardesty, W. R. Poller, R. Shah, and L. Wallace, "Changes in breast cancer detection and mammography recall rates after the introduction of a computer-aided detection system," J. Natl. Cancer Inst. **96**, 185–190 (2004).
[12]S. A. Feig, E. A. Sickles, W. P. Evans, and M. N. Linver, "Re. Changes in breast cancer detection and mammography recall rates after the introduction of a computer-aided detection system," J. Natl. Cancer Inst. **96**, 1260–1261 (2004).
[13]T. E. Cupples, "Impact of computer-aided detection (CAD) in a regional screening mammography program," Radiology **221**(P), 520 (2001).
[14]S. L. Ng and W. F. Bischof, "Automated detection and classification of breast tumors," Comput. Biomed. Res. **25**, 218–237 (1992).
[15]N. Petrick, H. P. Chan, D. Wei, B. Sahiner, M. A. Helvie, and D. D. Adler, "Automated detection of breast masses on mammograms using adaptive contrast enhancement and texture classification," Med. Phys. **23**, 1685–1696 (1996).
[16]B. Zheng, Y. H. Chang, and D. Gur, "Computerized detection of masses in digitized mammograms using single-image segmentation and a multilayer topographic feature analysis," Acad. Radiol. **2**, 959–966 (1995).
[17]N. Karssemeijer and G. te Brake, "Detection of stellate distortions in mammograms," IEEE Trans. Med. Imaging **15**, 611–619 (1996).
[18]H. Kobatake and Y. Yoshinaga, "Detection of spicules on mammogram based on skeleton analysis," IEEE Trans. Med. Imaging **15**, 235–245 (1996).
[19]Z. M. Huo, M. L. Giger, C. J. Vyborny, D. E. Wolverton, R. A. Schmidt, and K. Doi, "Automated computerized classification of malignant and benign masses on digitized mammograms," Acad. Radiol. **5**, 155–168 (1998).
[20]W. Qian, L. H. Li, and L. P. Clarke, "Image feature extraction for mass detection in digital mammography: Influence of wavelet analysis," Med. Phys. **26**, 402–408 (1999).
[21]N. Petrick, H. P. Chan, B. Sahiner, M. A. Helvie, S. Paquerault, and L. M. Hadjiiski, "Breast cancer detection: Evalution of a mass detection algorithm for computer-aided diagnosis: Experience in 263 patients," Radiology **224**, 217–224 (2002).
[22]S. Paquerault, N. Petrick, H. P. Chan, B. Sahiner, and M. A. Helvie, "Improvement of computerized mass detection on mammograms: Fusion

of two-view information," Med. Phys. **29**, 238–247 (2002).
[23]B. Zheng, W. F. Good, D. R. Armfield, C. Cohen, T. Hertzberg, J. H. Sumkin, and D. Gur, "Performance change of mammographic CAD schemes optimized with most-recent and prior image databases," Acad. Radiol. **10**, 283–288 (2003).
[24]J. Wei, B. Sahiner, L. M. Hadjiiski, H. P. Chan, N. Petrick, M. A. Helvie, M. A. Roubidoux, J. Ge, and C. Zhou, "Computer aided detection of breast masses on full field digital mammograms," Med. Phys. **32**, 2827–2838 (2005).
[25]N. Petrick, H. P. Chan, B. Sahiner, M. A. Helvie, and S. Paquerault, "Evaluation of an automated computer-aided diagnosis system for the detection of masses on prior mammograms," San Diego, 2000.
[26]G. M. Te Brake, N. Karssemeijer, and J. Hendriks, "Automated detection of breast carcinomas not detected in a screening program," Radiology **207**, 465–471 (1998).
[27]D. M. Ikeda, R. L. Birdwell, K. F. O'Shaughnessy, E. A. Sickles, and R. J. Brenner, "Computer-aided detection output on 172 subtle findings on normal mammograms previously obtained in women with breast cancer detected at follow-up screening mammography," Radiology **230**, 811–819 (2004).
[28]D. Gur, J. S. Stalder, L. A. Hardesty, B. Zheng, J. H. Sumkin, D. M. Chough, B. E. Shindel, and H. E. Rockette, "Computer-aided detection performance in mammographic examination of masses: Assessment," Radiology **233**, 418–423 (2004).
[29]H. El-Shishini, M. S. Abdel-mottaleb, M. El-Raey, and A. Shoukry, "A multistage algorithm for fast classification of patterns," Pattern Recogn. Lett. **10**, 211–215 (1989).
[30]J. Y. Zhou and T. Pavlidis, "Discrimination of characters by a multi-stage recognition process," Pattern Recogn. **27**, 1539–1549 (1994).
[31]L. K. Hansen and P. Salamon, "Neural network ensembles," IEEE Trans. Pattern Anal. Mach. Intell. **12**, 993–1001 (1990).
[32]G. Rogova, "Combining the results of several neural network classifiers," Neural Networks **7**, 777–781 (1994).
[33]S. Hashem and B. Schmeiser, "Improving model accuracy using optimal linear combinations of trained neural networks," IEEE Trans Neural Networks **6**, 792–794 (1995).
[34]T. K. Ho, J. J. Hull, and S. N. Srihari, "Decision combination in multiple classifier systems," IEEE Trans. Pattern Anal. Mach. Intell. **16**, 66–75 (1994).
[35]D. A. Bell, J. W. Guan, and Y. Bi, "On combining classifier mass functions for text categorization," IEEE Trans. Knowl. Data Eng. **17**, 1307–1319 (2005).
[36]J. Cao, M. Ahmadi, and M. Shridhar, "Recognition of handwritten numerals with multiple feature and multistage classifier," Pattern Recogn. **28**, 153–160 (1995).
[37]A. Al-Ani and M. Deriche, "A new technique for combining multiple classifiers using the Dempster-Shafer theory of evidence," J. Artif. Intell. Res. **17**, 333–361 (2002).
[38]J. Wei, B. Sahiner, L. M. Hadjiiski, H. P. Chan, M. A. Helvie, and M. A. Roubidoux, "A dual computer-aided detection (CAD) system for improvement of mass detection on mammograms," RSNA Program Book 2004, p. 491.
[39]J. Wei, B. Sahiner, L. M. Hadjiiski, H. P. Chan, M. A. Helvie, M. A. Roubidoux, N. Petrick, C. Zhou, and J. Ge, "Computer aided detection of breast masses on mammograms: Performance improvement using a dual system," Proc. SPIE **5747**, 9–15 (2005).
[40]J. Wei, B. Sahiner, L. M. Hadjiiski, H. P. Chan, N. Petrick, M. A. Helvie, C. Zhou, and Z. Ge, "Computer aided detection of breast masses on full-field digital mammograms: False positive reduction using gradient field analysis," Proc. SPIE **5370**, 992–998 (2004).
[41]N. Petrick, H. P. Chan, B. Sahiner, and M. A. Helvie, "Combined adaptive enhancement and region-growing segmentation of breast masses on digitized mammograms," Med. Phys. **26**, 1642–1654 (1999).
[42]D. Wei, H. P. Chan, N. Petrick, B. Sahiner, M. A. Helvie, D. D. Adler, and M. M. Goodsitt, "False-positive reduction technique for detection of masses on digital mammograms: Global and local multiresolution texture analysis," Med. Phys. **24**, 903–914 (1997).
[43]M. M. Galloway, "Texture classification using gray level run lengths," Comput. Graph. Image Process. **4**, 172–179 (1975).
[44]M. J. Norusis, *SPSS for Windows Release 6 Professional Statistics* (SPSS, Chicago, IL, 1993).
[45]L. M. Hadjiiski, B. Sahiner, H. P. Chan, N. Petrick, M. A. Helvie, and M. N. Gurcan, "Analysis of temporal change of mammographic features: Computer-aided classification of malignant and benign breast masses," Med. Phys. **28**, 2309–2317 (2001).
[46]C. E. Metz, B. A. Herman, and J. H. Shen, "Maximum-likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data," Stat. Med. **17**, 1033–1053 (1998).
[47]B. Sahiner, H. P. Chan, N. Petrick, R. F. Wagner, and L. M. Hadjiiski, "Feature selection and classifier performance in computer-aided diagnosis: The effect of finite sample size.," Med. Phys. **27**, 1509–1522 (2000).
[48]C. M. Bishop, *Neural Networks for Pattern Recognition* (Clarendon, Oxford, 1995).
[49]J. A. Freeman and D. M. Skapura, *Neural Networks-Algorithms, Applications, and Programming Techniques* (Addison-Wesley, Reading, 1991).
[50]D. P. Chakraborty and L. H. L. Winter, "Free-response methodology: Alternate analysis and a new observer-performance experiment," Radiology **174**, 873–881 (1990).
[51]D. P. Chakraborty and K. S. Berbaum, "Observer studies involving detection and localization: Modeling, analysis, and validation," Med. Phys. **31**, 2313–2330 (2004).
[52]K. F. O'Shaughnessy, R. A. Castellino, S. L. Muller, and K. Benali, "Computer-aided detection (CAD) on 90 biopsy-proven breast cancer cases acquired on a full-field digital mammography (FFDM) system," Radiology **221**(P), 471 (2001).
[53]L. J. Warren Burhenne, S. A. Wood, C. J. D'Orsi, S. A. Feig, D. B. Kopans, K. F. O'Shaughnessy, E. A. Sickles, L. Tabar, C. J. Vyborny, and R. A. Castellino, "Potential contribution of computer-aided detection to the sensitivity of screening mammography," Radiology **215**, 554–562 (2000).
[54]R. E. Brem, J. W. Hoffmeister, J. A. Rapelyea, G. Zisman, K. Mohtashemi, G. Jindal, M. P. DiSimio, and S. K. Rogers, "Impact of breast density on computer-aided detection for breast cancer," AJR, Am. J. Roentgenol. **184**, 439–444 (2005).

Computer-Aided Detection Systems for Breast Masses: Comparison of Performances on Full-Field Digital Mammograms and Digitized Screen-Film Mammograms[1]

Jun Wei, PhD, Lubomir M. Hadjiiski, PhD, Berkman Sahiner, PhD, Heang-Ping Chan, PhD, Jun Ge, PhD
Marilyn A. Roubidoux, MD, Mark A. Helvie, MD, Chuan Zhou, PhD, Yi-Ta Wu, PhD
Chintana Paramagul, MD, Yiheng Zhang, PhD

Rationale and Objectives. To compare the performance of computer aided detection (CAD) systems on pairs of full-field digital mammogram (FFDM) and screen-film mammogram (SFM) obtained from the same patients.

Materials and Methods. Our CAD systems on both modalities have similar architectures that consist of five steps. For FFDMs, the input raw image is first log-transformed and enhanced by a multiresolution preprocessing scheme. For digitized SFMs, the input image is smoothed and subsampled to a pixel size of 100 μm $\times$ 100 μm. For both CAD systems, the mammogram after preprocessing undergoes a gradient field analysis followed by clustering-based region growing to identify suspicious breast structures. Each of these structures is refined in a local segmentation process. Morphologic and texture features are then extracted from each detected structure, and trained rule-based and linear discriminant analysis classifiers are used to differentiate masses from normal tissues. Two datasets, one with masses and the other without masses, were collected. The mass dataset contained 131 cases with 131 biopsy proven masses, of which 27 were malignant and 104 benign. The true locations of the masses were identified by an experienced Mammography Quality Standards Act (MQSA) radiologist. The no-mass data set contained 98 cases. The time interval between the FFDM and the corresponding SFM was 0 to 118 days.

Results. Our CAD system achieved case-based sensitivities of 70%, 80%, and 90% at 0.9, 1.5, and 2.6 false positive (FP) marks/image, respectively, on FFDMs, and the same sensitivities at 1.0, 1.4, and 2.6 FP marks/image, respectively, on SFMs.

Conclusions. The difference in the performances of our FFDM and SFM CAD systems did not achieve statistical significance.

Key Words. Computer-aided detection; mass detection; full-field digital mammogram (FFDM); screen-film mammogram (SFM); free-response receiver operating characteristic (FROC).

Acad Radiol 2007; 14:659–669

[1] From the Department of Radiology, University of Michigan, CGC B2103, 1500 E. Medical Center Drive, Ann Arbor, MI 48109 (J.W., L.M.H., B.S., H.-P.C., J.G., M.A.R., M.A.H., C.Z., Y.-T.W., C.P., Y.Z.). Received December 15, 2006; accepted February 13, 2007. This work is supported by US Army Medical Research and Materiel Command grants W81XWH-1-04-1-0475 and DAMD 17-02-1-0214, and USPHS grant CA95153. The content of this article does not necessarily reflect the position of the government and no official endorsement of any equipment and product of any companies mentioned should be inferred. The authors are grateful to Charles E. Metz, PhD, for the LABROC program and to Dev Chakraborty, PhD, for the JAFROC program. **Address correspondence to:** J.W. e-mail: jvwei@umich.edu.

doi:10.1016/j.acra.2007.02.017

Full-field digital mammography (FFDM) and screen-film mammography (SFM) are two available methods for breast cancer screening in clinical practice. FFDM detectors provide higher detective quantum efficiency (DQE) and signal-to-noise ratio (SNR), wider dynamic range, and higher contrast sensitivity than SFM. FFDM may alleviate some of the limitations of SFM, especially in breasts with dense fibroglandular tissue (1). In the last few years, several FFDM systems became commercially available because of the potential of digital imaging to improve breast cancer detection.

Several clinical trials have been conducted to compare radiologists' interpretation on FFDMs and SFMs. Lewin et al (2,3) conducted a clinical study to compare FFDMs and SFMs for the detection of breast cancer in 6,737 examinations of women 40 years of age and older collected from two institutions. Forty-two cancers were detected within this population. The difference in cancer detection was not statistically significant ($P > .1$) between FFDMs and SFMs. FFDMs resulted in fewer recalls than did SFM, which was statistically significant ($P < .001$). Another clinical trial (4) aiming at collecting data for US Food and Drug Administration approval included SFMs and FFDMs of 676 women who were scheduled to undergo breast biopsy. The average area under the receiver operating characteristic (ROC) curve, the sensitivity and the specificity were 0.715, 0.66 and 0.67 for printed FFDM and 0.765, 0.74, 0.60 for SFM, respectively. However, none of these differences achieved statistical significance. Skaane et al (5–7) has conducted several clinical studies to compare SFM and FFDM with soft-copy interpretation for reader performance in detection and classification of breast lesions. According to their findings, there was no significant difference between FFDM and SFM either in detection or in classification. A recent study by Pisano et al (1) collected a total of 49,528 patients at 33 sites in the United States and Canada. Mammograms were interpreted independently by two radiologists. The overall diagnostic accuracy of FFDMs and SFMs for breast cancers was similar. However, FFDM was more accurate in women younger than age 50 years, women with radiographically dense breasts, and premenopausal or perimenopausal women.

Studies indicate that radiologists do not detect all carcinomas that are visible on retrospective analyses of the images (8–14). Computer-aided diagnosis (CAD) is considered to be one of the promising approaches that may improve the sensitivity of mammography (15,16). Most of the mammographic CAD systems developed so far are based on digitized SFMs. Li et al (17) attempted to adapt their CAD system developed on SFMs for detection of masses on FFDMs by standardizing the FFDMs. Their preliminary results on a small data set (training on 36 normal and 24 mass cases, testing on 24 normal and 10 mass cases) showed 60% sensitivity at 2.47 false positives (FPs)/image. Several commercial CAD systems reported comparable performance on FFDMs and SFMs. However, their study was not reported in peer-reviewed journals, so that the dataset and algorithm are unknown. So far, there are no studies on comparison of breast mass detection between FFDMs and SFMs from the same patients by using CAD system. We have developed a CAD system for mass detection on SFMs (18,19) and are adapting the system to FFDMs. Our preliminary study with 65 patients was reported previously (20). In this study, we compared the performance of the two CAD systems on case-matched pairs of FFDMs and SFMs.

MATERIALS AND METHODS

Materials

Our study group consisted of patients with breast lesions that were categorized suspicious and recommended for biopsy. The patients had either FFDM or SFM for their clinical exams. Institutional review board approval and patient informed consent were obtained to acquire corresponding mammograms of the breast to be biopsied using the other modality. Therefore, the corresponding FFDM and SFM were available only from one breast for each patient. The time interval between the SFM and the FFDM ranged from 0 to 118 days. The dataset consisted of 229 patients aged 30–86 with a mean age of 55 ± 11 years. All cases have two mammographic views, the craniocaudal view and the mediolateral oblique view or the lateral view, yielding a total of 458 FFDMs and 458 corresponding SFMs. The SFMs were acquired with MinR2000 screen-film systems (Eastman Kodak, Rochester, NY) and digitized with a LUMISCAN 85 laser film scanner (Lumisys, Los Altos, CA) at a pixel resolution of 50 μm × 50 μm and 4096 gray levels. The digitizer was calibrated so that gray-level values were linearly proportional to the optical density in the range of 0–4, with a slope of 0.001 per pixel value. The digitizer output was linearly converted so that a large pixel value corresponded to a low optical density. FFDMs were acquired with a GE Senographe 2000D system (GE Medical Systems, Milwaukee, WI). The GE system has a CsI

Table 1
Description of Cases in the Mass Datasets and Subsets for Training and Testing in the Twofold Cross-Validation Scheme

	Mass Set		Mass Subset 1		Mass Subset 2	
	FFDM	SFM	FFDM	SFM	FFDM	SFM
Total number of cases	131	131	65	65	66	66
Total number of images	262	262	130	130	132	132
Number of visible masses (by case)	131	130	65	65	66	65
Number of masses only visible on one view	8	9	5	5	3	4
Number of visible masses (by image)	254	251	125	125	129	126
Number of visible malignant masses	27	27	12	12	15	15
Number of visible benign masses	104	103	53	53	51	50

FFDM: full-field digital mammogram; SFM: screen-film mammogram.

phosphor/a:Si active matrix flat panel digital detector with a pixel size of 100 μm $\times$ 100 μm and 14 bits per pixel. The raw FFDMs were used as the input of our CAD system.

The dataset included 131 cases containing masses and 98 cases containing microcalcifications without a visible mass, as determined with visual inspection by an experienced radiologist. The 131 cases will be referred to as the mass dataset and the 98 cases as the "no-mass" data set in the following discussion. The no-mass cases were considered as "normal" with respect to masses and were used to estimate the FP mark rates of the CAD systems during testing. The mass dataset contained 131 biopsy proved masses, of which 27 were malignant and 104 benign. By examining all available information, including the diagnostic mammograms and reports, the true locations of the masses were identified by an experienced Mammography Quality Standards Act (MQSA) radiologist. In these 131 mass cases, 1 mass can be seen only on FFDMs, 7 masses can be seen on only one view on both FFDMs and SFMs, and 3 masses can be seen on only one view on either FFDMs (1 mass) or SFMs (2 masses). There were therefore 131 visible masses on FFDMs and 130 visible masses on SFMs if the masses were counted by case. There were 254 visible and 8 invisible masses on FFDMs and 251 visible and 11 invisible masses on SFMs if the masses were counted independently by mammographic view. The number of images and masses in the mass dataset are described in Table 1. Figure 1 shows an example with a 7-mm malignant mass. The size of a mass was estimated as its longest diameter seen on the mammograms. The visibility of the masses was rated by the experienced radiologist on a 10-point scale, with 1 representing the most visible masses and 10 the most difficult case relative to the cases seen in clinical practice. Figures 2 and 3 show the histograms of mass sizes and visibility, respectively, for the mass set. The mass size ranged from 3 to 30 mm (mean: 12.5 $\pm$ 4.9 mm on FFDMs and 12.6 $\pm$ 4.9 mm on SFMs) and the visibility ratings extended over the entire range. Figure 4 shows the breast density in terms of BI-RADS category as estimated by the radiologist for the FFDM and SFM datasets.

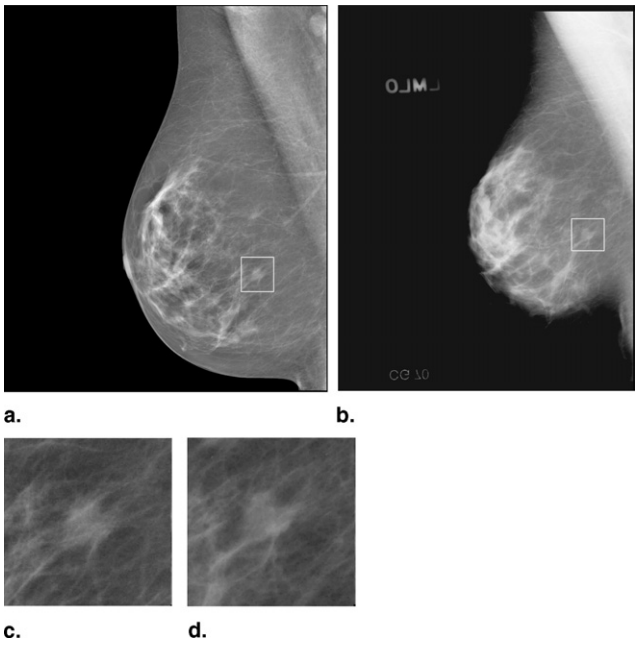


Figure 1. An example of mammograms with a region of interest (ROI) containing a malignant mass with a size of 7 mm. **(a)** Processed full-field digital mammogram (FFDM) by using the Laplacian pyramid multiscale method, **(b)** digitized screen-film mammogram (SFM), **(c)** magnified ROI on FFDM, and **(d)** magnified ROI on SFM. The SFM is displayed with the same resolution as that of the FFDM. The apparently smaller breast size on SFM is mainly caused by the very dark breast periphery region on the SFM that cannot be seen on the printed page.

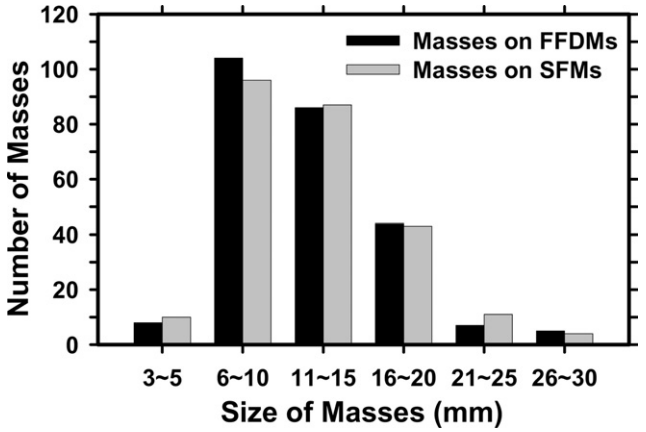


Figure 2. Histogram of the sizes for 254 masses on full-field digital mammograms (FFDMs) and 251 masses on the screen-film mammograms (SFMs) in our dataset. Mass sizes are measured as the longest dimension of the mass by an experienced Mammography Quality Standards Act (MQSA) radiologist. The size of the masses in the dataset ranged from 3 to 30 mm (mean: 12.5 ± 4.9 mm on FFDMs and 12.6 ± 4.9 mm on SFMs).

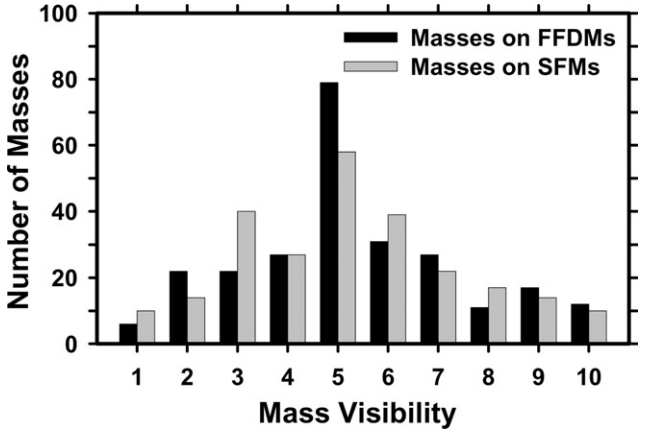


Figure 3. Histogram of the visibility of the 254 masses seen on full-field digital mammograms and 251 masses seen on screen-film mammograms in our dataset. The visibility is evaluated on a 10-point rating scale, with 1 representing the most visible masses and 10 the most difficult case relative the cases seen in their clinical practice. Each mass on a mammogram is rated independently by an experienced MQSA radiologist.

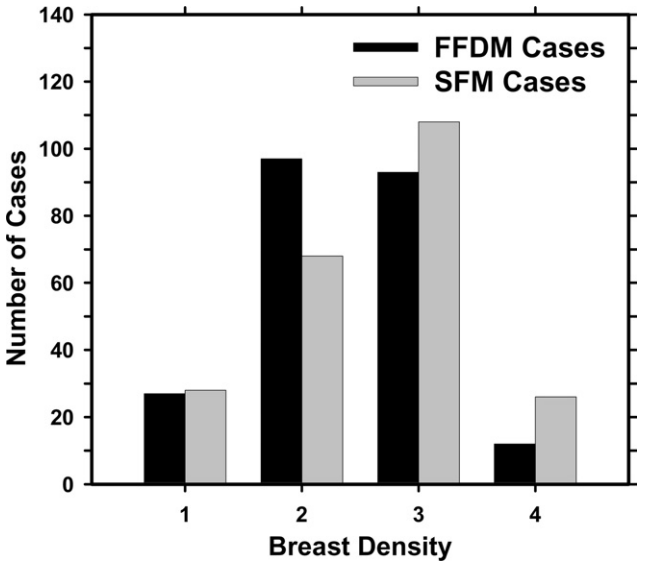


Figure 4. Distribution of the breast density for the 229 cases in terms of BI-RADS category estimated by an MQSA radiologist.

METHODS

CAD System

The major steps in the mass detection systems on FFDMs and SFMs are similar, but the feature spaces and classifiers for FP reduction in each system were designed separately to suit the characteristics of FFDMs and SFMs. The two systems are therefore described together, but the differences will be pointed out whenever applicable. Each single CAD system consists of five processing steps: 1) preprocessing, 2) prescreening of mass candidates, 3) segmentation of suspicious objects, 4) feature extraction and analysis, and 5) FP reduction by classification of normal tissue structures and masses.

FFDMs are generally preprocessed with proprietary methods by the manufacturer of the FFDM system before being displayed to readers. The image preprocessing method used depends on the manufacturer of the FFDM system. To develop a CAD system that is less dependent on the FFDM manufacturer's proprietary preprocessing methods, we use the raw FFDM as input to our CAD system. We have previously developed a multiscale preprocessing scheme for image enhancement (21). In brief, the raw mammogram is first segmented automatically into the background and the breast region. A logarithmic transform is applied to the image which is then scaled to 12-bit. The Laplacian pyramid method (21,22) is used to decompose the transformed breast image into multiscales. A nonlinear weight function based on the pixel gray level from each of the low-pass components is designed to enhance the high-pass components. The processed image is reconstructed by summing the weighted components.

For SFMs, the full resolution digitized mammograms are smoothed with a 2×2 box filter and subsampled by a factor of 2, resulting in images having a pixel size of $100\ \mu m \times 100\ \mu m$. These images are used as input to the CAD system.

After preprocessing, a two-stage gradient field analysis method (21,23) is used to identify the mass candidates for

either FFDMs or SFMs. In brief, a gradient field analysis is employed in the first stage to identify potential mass candidates based on high values of the initial gradient field. Each potential mass candidate is segmented by a region growing technique. The shape and the gray-level information of the segmented object allow adaptive refinement of the gradient field analysis in the second stage. Locations of high radial gradient convergence are then labeled as mass candidates. These suspicious objects are segmented with a k-means clustering method (24). First, a 256×256 pixel region of interest (ROI) centered at the high gradient point is background-corrected (25) and weighted by a Gaussian function with $\sigma = 256$ pixels. K-means clustering using the pixel values in a background-corrected image and a Sobel filtered image as features is then used to segment the object.

For each suspicious object, eleven morphological features (18) are extracted. A rule-based classifier removes the detected structures that are substantially different from breast masses. Global and local multiresolution texture analyses (26) are performed in each ROI by using the spatial gray-level dependence (SGLD) matrices. Thirteen SGLD texture measures are used. Global texture features are extracted from the entire ROI for two scales, seven distances, and two angles. Local texture features are extracted from the local region containing the detected object and the peripheral regions within each ROI for two scales, four distances, and two angles. Therefore, a total of 364 features and 208 features, respectively, are extracted from global and local texture analysis. The feature space for final classification is the combination of morphologic features and SGLD texture features. Finally, linear discriminant analysis (LDA) is used to classify masses from normal tissue in the feature space. The discriminant scores are ranked for each mammogram, and any object with a discriminant score that ranks lower than three is eliminated.

Training and Test CAD System

Twofold cross-validation was used for training and testing our CAD system for FFDMs. We randomly separated the mass datasets by case into two independent subsets: subset 1 with 65 cases and subset 2 with 66 cases. The numbers of masses by image and by case for the FFDM and SFM subsets are shown in Table 1. The training included selection of proper parameters and features for the classifier in the CAD system. After the training with one mass subset was completed, the parameters and features were fixed for testing with the other mass subset. The training and test mass subsets were switched and the training and test processes were repeated. The trained CAD systems were also applied to the no-mass data set, which was not used during training, to estimate the FP rate in screening mammograms.

During training, feature selection with stepwise LDA was applied to obtain the best feature subset and reduce the dimensionality of the feature space to design an effective classifier. The detailed procedure has been described elsewhere (21,27,28). Briefly, at each step one feature was entered or removed from the feature pool by analyzing its effect on the selection criterion, which was chosen to be the Wilks' lambda in this study. Because the appropriate threshold values for feature entry, feature elimination, and tolerance of feature correlation were unknown, we used an automated simplex optimization method to search for the best combination of thresholds in the parameter space. The simplex algorithm used a leave-one-case-out resampling method within the training subset to select features and estimate the weights for the LDA classifier. To have a figure of merit to guide feature selection, the test discriminant scores from the left-out cases were analyzed using ROC methodology (29). The accuracy for classification of masses and FPs was evaluated as the area under the ROC curve, A_z, for the test cases. In this approach, feature selection was performed without the left-out case so that the test performance would be less optimistically biased (30). However, the selected feature set in each leave-one-case-out cycle could be slightly different because every cycle had one training case different from the other cycles. To obtain a single trained classifier to apply to the cross-validation test subset, a final stepwise feature selection was performed with the best combination of thresholds, found in the simplex optimization procedure, on the entire training subset to obtain the final set of features and estimate the weights of the LDA. Note that the entire process of feature selection and classifier weight estimation was performed within the training subset. The LDA classifier with the selected feature set was then fixed and applied to the cross-validation test subset. The training and testing processes were performed independently for the twofold cross-validation sets.

Because we already trained our CAD system for SFMs with a large dataset in a previous study (19), we used the trained system without retraining the parameters in this study. For testing, we divided the SFMs into two test datasets that followed the same case grouping as that for FFDMs. The test cases in each subset did not overlap

with any training cases used for training the SFM CAD system in the previous study.

Evaluation Methods

We used a free-response ROC (FROC) method (31) to assess the overall performance of the CAD scheme on this image set. A FROC curve is obtained by plotting the mass detection sensitivity as a function of FP marks per image as the decision threshold on the LDA classifier scores varies.

The detected individual objects were compared with the "true" mass locations marked by the experienced radiologist, as described previously. A detected object was labeled as true positive (TP) if the overlap between the bounding box of the detected object and the bounding box of the true mass relative to the larger of the two bounding boxes was over 25%. Otherwise, it would be labeled as FP. The 25% threshold was selected as described in our previous study (18).

FROC curves were presented on a per-image and a per-case basis. For image-based FROC analysis, the mass on each mammogram was considered an independent true object; the sensitivity was thus calculated relative to the number of masses by image on each subset of FFDMs or SFMs (Table 1). For case-based FROC analysis, the same mass imaged on the two-view mammograms was considered to be one true object and detection of either or both masses on the two views was considered to be a TP detection; the sensitivity was thus calculated relative to the number of masses by case on each subset of FFDMs or SFMs (Table 1). The test FROC curve for a given mass subset was estimated by counting the detected masses on the test mass subset for the sensitivity. The FP marker rate was estimated in two ways: one from FPs detected in the same test mass subsets, the other from FPs detected in the no-mass dataset. For the latter, we applied the trained CAD system to the entire no-mass dataset. The average number of FP marks per image produced by the CAD system at a given sensitivity was estimated by counting the detected objects in these cases at the corresponding decision threshold. Because we used twofold cross-validation method for training and testing, we obtained two test FROC curves, one for each test subset, for each of the modalities. To summarize the results for comparison, an average test FROC curve was derived by averaging the FP rates at the same sensitivity along the FROC curves of the two corresponding test subsets.

To compare the performance of our CAD system for FFDMs and SFMs statistically, we applied the alternative free-response ROC (AFROC) method and the jackknife free-response ROC (JAFROC) method developed by Chakraborty et al (32,33) to the pairs of FROC curves. In the AFROC method, the FROC data are first transformed by counting the number of false-positive images instead of the FPs per image. The LDA score of a false-positive image is determined by the highest score FP object on the image regardless of how many lower scores FP objects are made on the same image. The ROCKIT curve fitting software and statistical significance tests for ROC analysis developed by Metz et al (29) can then be used to analyze the AFROC data.

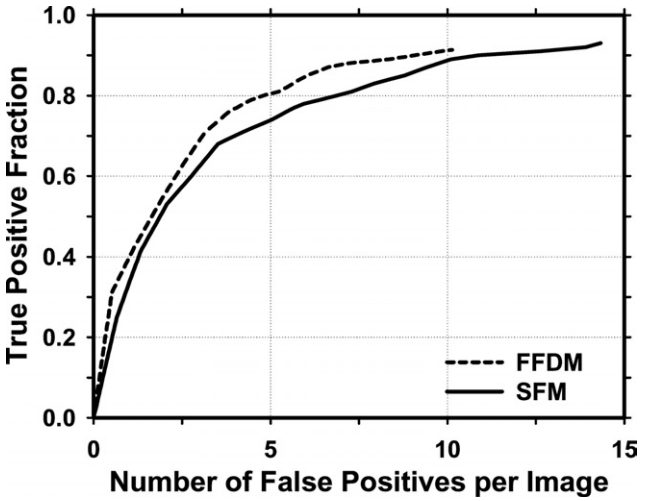


Figure 5. Comparison of free-response receiver operating characteristic (FROC) curves on full-field digital mammograms and screen-film mammograms during the prescreening stage. The FROC curves were generated by varying the number of detected suspicious objects per image based on the ranking of the local maxima on gradient field images. The FP rate was estimated from the mammograms with masses.

RESULTS

For simplicity, we combined the detection results on the two test subsets from the twofold cross-validation process in the following discussion. The prescreening stage detected 91.3% (232/254) of the masses with an average of 10.13 (2,655/262) FPs /image on FFDMs and 93.2% (234/251) with an average of 14.43 (3,781/262) FPs/image on SFMs. Figure 5 compares the FROC curves on FFDMs and SFMs during the prescreening stage. The FROC curves were generated by varying the number of detected suspicious objects per image based on the ranking of local maxima on the gradient field images.

We used two steps for FP reduction for both CAD systems. The first step was the rule-based classification based on morphologic features. After this step, there were 2,572 mass candidates (9.8 objects/image) on FFDMs and 3,654 mass candidates (13.9 objects/image) on SFMs without additional FNs for the test sets of 262 images. The second step was the LDA classification. A total of 16 (4 global texture features, 7 local texture features, and 5 morphologic features) and 12 (4 global texture features, 4 local texture features, and 4 morphologic features) features, respectively, were selected from the two independent training subsets for FFDMs. The feature set for SFMs contained a total of 21 features (11 global texture features, 7 local texture features, and 3 morphologic features), as obtained from previous training.

Figure 6 shows the comparison of the average test FROC curves of the CAD systems for FFDMs and SFMs. The FFDM CAD system achieved a case-based sensitivity of 70%, 80%, and 90% at 0.67, 1.15, and 1.93 FPs/image, respectively, compared with 0.75, 1.06, and 1.86 FPs/image for the SFM CAD system. Because two trained CAD systems were obtained for the FFDMs from the cross-validation training, we applied each of the trained systems to the no-mass data set for FROC analysis, and estimated the number of FP marks per image on the no-mass cases at each decision threshold. For each trained CAD system, the sensitivity was estimated from the detected masses on the test mass subset and plotted against the FP rate estimated from the no-mass set. Figure 7 shows the average FROC curves for FFDMs and SFMs, similar to those shown in Fig 6, except that the FP rates were estimated from the no-mass data set.

The comparison of the FROC curves for the FFDM and SFM CAD systems in terms of the area under the fitted AFROC curve (A_1) and the P values for both test mass subsets are summarized in Table 2. The differences in the A_1 values between the two modalities did not achieve statistical significance ($P > .05$). The fitted AFROC curves, however, did not fit very well to the transformed AFROC data, as discussed previously (21). For the JAFROC method, Chakraborty et al provided software to estimate the statistical significance of the difference between two FROC curves. The comparison of the figure-of-merit (FOM) and the P values is also summarized in Table 2. The differences in the FOMs between the FFDM and SFM CAD systems again did not achieve statistical significance ($P > .05$).

There were 27 malignant cases in the mass set. Figure 8 compares the average test FROC curves of the FFDM and SFM CAD systems for detection of malignant masses. The FP rate was estimated from the no-mass dataset. In this case, the FFDM CAD system achieved a case-based sensitivity of 70%, 80%, and 90% at 0.37, 0.73, and 1.31 FP marks/image, respectively, which were substantially better than the FP rates of 1.1, 1.6, and 2.0 FP marks/image for the SFM CAD system. However, the difference did not achieve statistical significance ($P > .05$).

A total of 105 FFDM cases and 134 SFM cases were identified as BI-RADS 3 and 4 categories by an MQSA

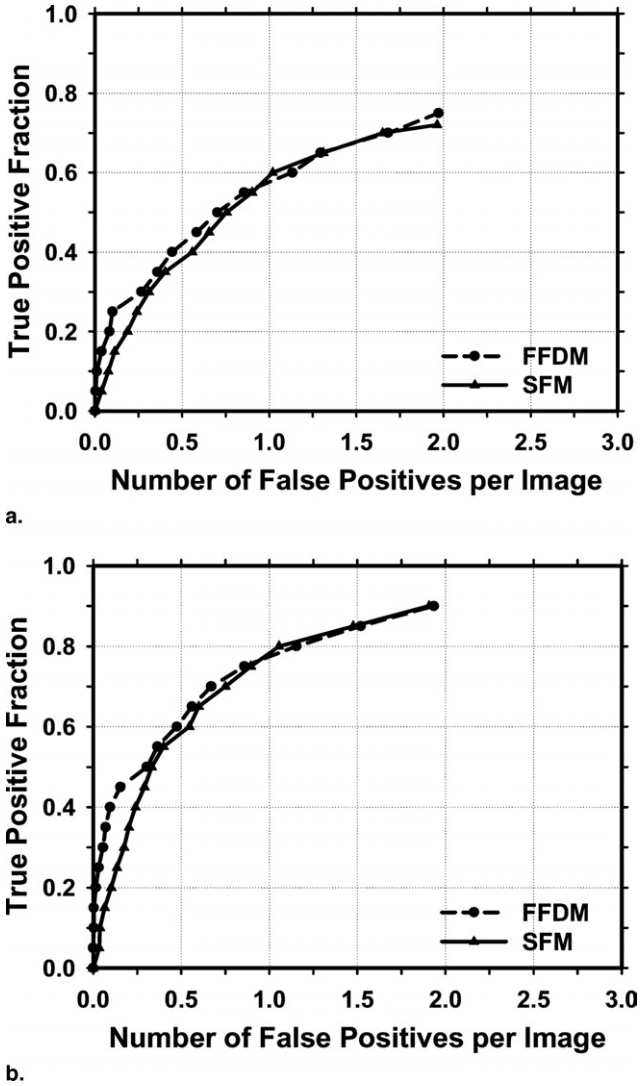
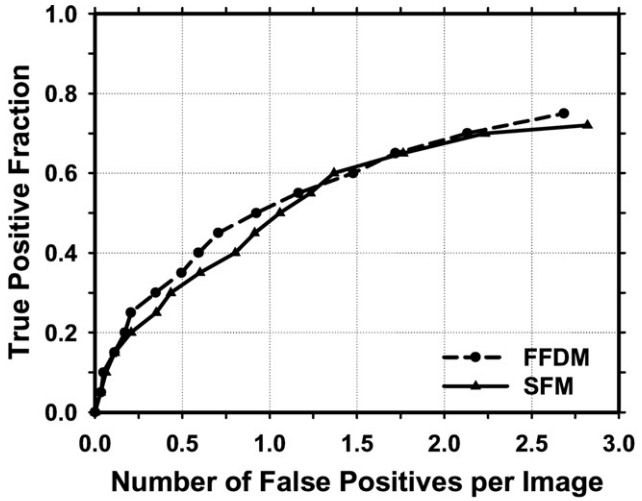
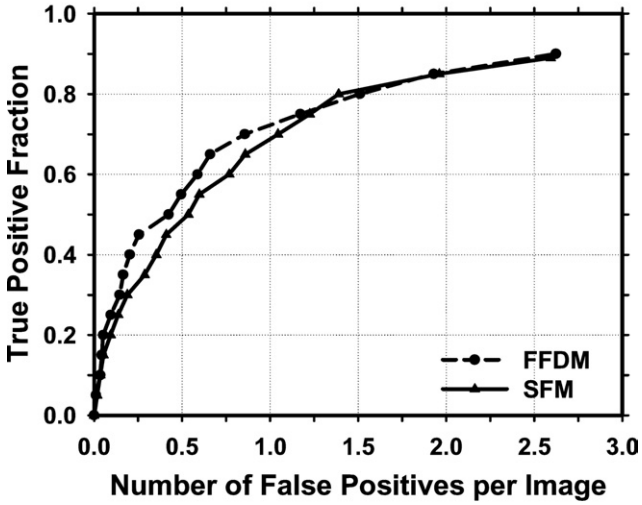


Figure 6. Comparison of the average test free-response receiver operating characteristic (FROC) curves obtained from averaging the FROC curves of the two independent mass subsets on full-field digital mammograms and screen-film mammograms. The FP rate was estimated from the mammograms with masses. **(a)** Image-based FROC curves and **(b)** case-based FROC curves.



a.



b.

Figure 7. Comparison of the average test free-response receiver operating characteristic (FROC) curves obtained from averaging the FROC curves of the two independent mass subsets on full-field digital mammograms and screen-film mammograms. The FP rate was estimated from the mammograms without masses. **(a)** Image-based FROC curves and **(b)** case-based FROC curves.

radiologist (Fig 4). Of these, 88 cases (56 mass cases and 32 no-mass cases) were in common. Figure 9 compares the average test FROC curves of the FFDM and SFM CAD systems for detection of masses only on this common subset of dense breasts. The FP rate was estimated from the 32 no-mass dense breasts. Although the FROC curve for the FFDMs appears to be slightly higher than that of the SFMs, the difference did not achieve statistical significance ($P > .05$).

DISCUSSION

CAD systems have been proven to be helpful as a second opinion to assist radiologists in interpretation of SFMs. Recently several studies have been conducted to compare FFDM with SFM in screening cohorts (1,4,5,34). These clinical trials arrived at different conclusions about the advantages or disadvantages of FFDM in comparison to conventional SFM systems. Some of the differences may be attributed to factors such as the mammographic equipment, the study design, the sample sizes, and the reader experience. It is also important to compare the performances of FFDM and SFM CAD systems. In our study, we compared the performance of the two systems on pairs of FFDM and SFM obtained from the same patients at close time intervals.

Several FFDM systems have been approved for clinical applications. Because digital detectors generally have a linear response to x-ray exposure, the raw pixel values are a linear function of the absorbed x-ray energy in the detector. To develop a CAD system that is less dependent on the FFDM manufacturer's proprietary preprocessing methods, we used the raw FFDM as input to our CAD system. Although the spatial resolution and noise properties of the images from different detectors were still different, the use of raw images already reduced one of the major differences between mammograms from different FFDM systems. For preprocessing of the raw FFDMs, we developed a multiresolution enhancement method. From our observation on the SFMs and the processed FFDMs, the breast tissue on SFMs appears to be denser than that on FFDMs (35). This may be attributed to the harder beam quality used and the Laplacian enhancement on FFDMs. In this study, 134 SFM cases were rated as BI-RADS 3 and 4 categories by an MQSA radiologist, whereas only 105 FFDM cases were rated as BI-RADS 3 and 4. When the FFDM and SFM CAD systems were applied to the small common subset (56 with masses and 32 without masses) of dense breasts rated as BI-RADS 3 and 4, there was no significant difference between their average test FROC curves (Fig 9).

The overall performances of the CAD systems for the two modalities did not demonstrate significant difference for comparisons in either the subsets or the entire dataset. One factor may be the substantially smaller number of training samples used for the FFDM CAD system than that for the SFM CAD system, which was trained with a set of 486 SFMs in a previous study (19). We have

Table 2
Estimation of the Statistical Significance of the Difference in the FROC Performances Between the FFDM and SFM CAD Systems

	A_1 (AFROC)				FOM (JAFROC)			
	All Cases		Malignant Cases		All Cases		Malignant Cases	
	Test Subset 1	Test Subset 2	Test Subset 1	Test Subset 2	Test Subset 1	Test Subset 2	Test Subset 1	Test Subset 2
FFDM	0.48	0.49	0.51	0.49	0.47	0.48	0.55	0.47
SFM	0.42	0.43	0.47	0.42	0.46	0.41	0.48	0.42
P values	.17	.16	.56	.23	.73	.33	.29	.59

The FROC curves with the FP marker rates obtained from the no-mass dataset were compared.

FROC, ; FFDM, full-field digital mammogram; SFM, screen-film mammogram; CAD, computed-aided detection; AFROC, alternative free-response receiver operating characteristic; FOM, figure-of-merit; JAFROC, jackknife free-response ROC.

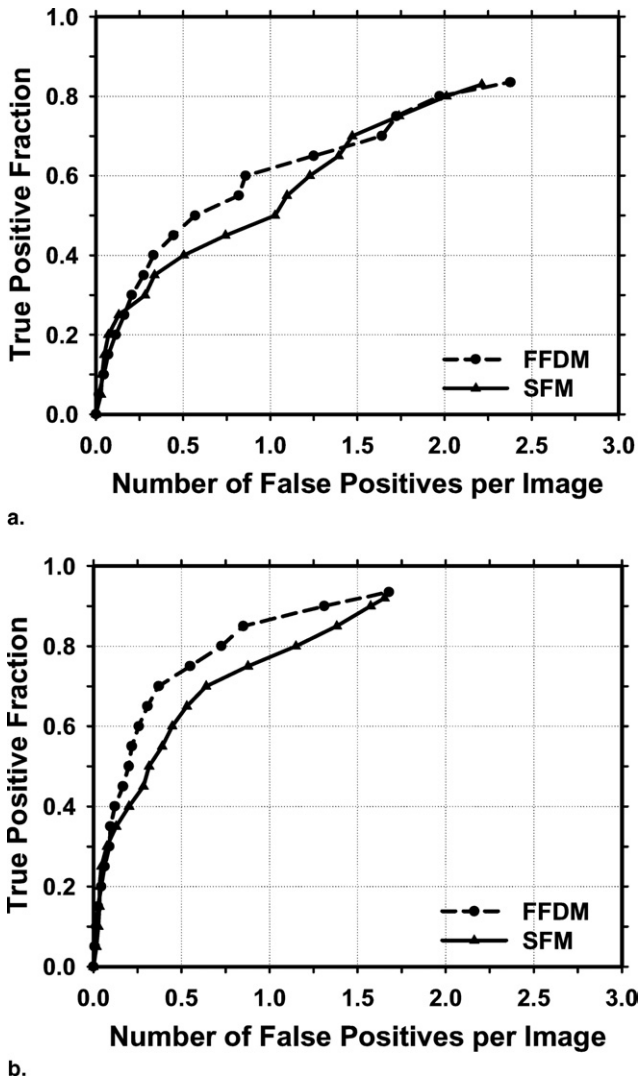


Figure 8. Comparison of the average test free-response receiver operating characteristic (FROC) curves of computed-aided detection systems on full-field digital mammograms and screen-film mammograms for mammograms with malignant masses. The FP rate was estimated from the mammograms without masses. **(a)** Image-based FROC curves and **(b)** case-based FROC curves.

shown previously that a classifier designed with a larger number of training samples will have better generalization to unknown test cases (36). Furthermore, because our CAD system was originally developed on SFMs, some of those techniques used may favor SFMs. If new techniques are designed to specifically suit the properties of FFDMs, the biases may be reduced. Further investigations are underway to improve the FFDM CAD system.

We used a twofold cross-validation method for training and testing of the CAD systems. Feature selection and classifier weight design were performed within the training subset and thus were independent of the test subset. Kupinski et al (37) showed that feature selection and classifier weight design using the same training set of a limited size will introduce additional optimistic bias to the training result and thus additional pessimistic bias to the test result. Under the constraint of a limited training set, the relative gain or loss in terms of bias if the training set is further split into two subsets for separate feature selection and classifier weight design in comparison to using the entire set of available training samples for both processes is still unknown. The relative efficiency of different resampling techniques in utilization of a limited dataset for classifier design with or without feature selection remains an important area of further studies. In screening mammography, the cancer rate is about 3–5 per 1,000. Most of the mammograms are normal. Therefore, some CAD researchers and users estimate the FP rate using normal mammograms (38–40) because it reflects how the CAD system performs in terms of specificity in a screening setting. Furthermore, for CAD systems that set a maximum number of detected objects at the output, estimating the number of FPs using images with lesions can potentially lead to an optimistic bias for the FROC curve because one of the detected objects will likely be the true

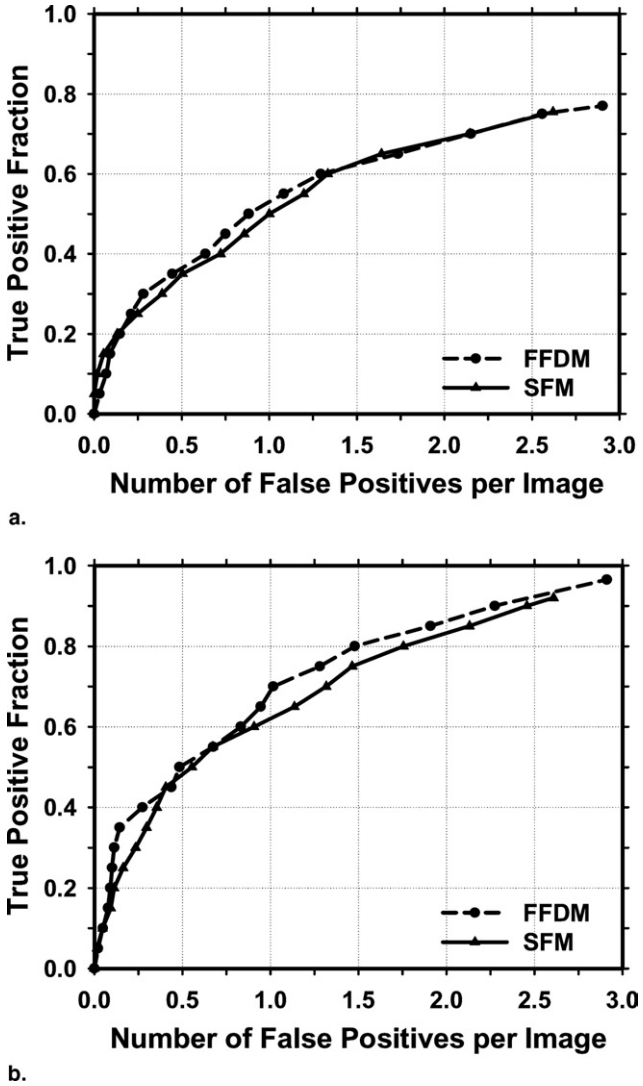


Figure 9. Comparison of the average test free-response receiver operating characteristic (FROC) curves of computed-aided detection systems on full-field digital mammograms and screen-film mammograms for the common subset of 56 dense breasts with masses rated as BI-RADS 3 and 4. The FP rate was estimated from 32 no-mass dense breasts that were also rated as BI-RADS 3 and 4. **(a)** Image-based FROC curves and **(b)** case-based FROC curves.

lesion. The FP rate can thus be underestimated by as much as 1 per image. In addition, the JAFROC analysis requires that the FP rates be estimated on normal images. We therefore reported the FP rates of our CAD systems on both mammograms with masses and without masses to facilitate comparison with other CAD systems in case investigators may evaluate their FP rates in either way.

Although we collected case-matched cases for comparing the performances of the CAD systems for FFDMs and SFMs, the images may not be exactly matched. Variations from positioning, compression force, and the difference in time between the two acquisitions would cause differences in the subtlety of the masses on the FFDMs and SFMs. However, assuming that the differences are random, both datasets would include images that have better or worse positioning, for example, than that on the other modality. The differences in the various factors would likely be averaged out over the entire dataset. We expect that they might not cause substantial bias in the comparison of the relative performances of the CAD systems for the two modalities.

For a CAD system, its performance for detecting malignant masses is more important than its performance for detecting all masses. We only have 27 malignant cases in this dataset. Although the FROC curves for detection of malignant masses (Fig 8) indicated that the FFDM CAD system had a higher sensitivity than that of the SFM CAD system, the differences in the A_1 and the FOM did not achieve statistical significance ($P > .05$) for either test subsets, as shown in Table 2. A large dataset is being collected for further comparison of the FFDM and SFM CAD systems for breast cancer cases.

Conclusion

We compared the performance of our CAD systems for detection of breast masses on case-matched FFDM images and SFM images. The two CAD systems used similar computer vision techniques but their preprocessing methods were different and the FP classifiers were separately trained to adapt to the image properties of each modality. From the comparison of FROC curves, it was found that the FFDM CAD system achieved higher detection sensitivity than the SFM CAD system at the same FP rates for malignant cases. However, the performances of our FFDM and SFM CAD systems for the entire data set were similar. The differences between the two modalities were not statistically significant with both AFROC and JAFROC methods for either the entire dataset or the malignant cases alone. Further study is under way to collect a larger dataset and to improve the performances of both systems.

REFERENCES

1. Pisano ED, Gasonis C, Hendrick E, et al. Diagnostic performance of digital versus film mammography for breast-cancer screening. N Engl J Med 2005; 353:1773–1783.
2. Lewin JM, Hendrick RE, D'Orsl CJ, et al. Comparison of full-field digital mammography with screen-film mammography for cancer detection: results of 4,945 paired examinations. Radiology 2001; 218:873–880.

3. Lewin JM, D'Orsi CJ, Hendrick RE, et al. Clinical comparison of full-field digital mammography and scree-film mammography for detection of breast cancer. AJR Am J Roentgenol 2002; 179:671–677.
4. Cole E, Pisano E, Brown M, et al. Diagnostic accuracy of Fischer SenoScan digital mammography versus screen-film mammography in a diagnostic mammography population. Acad Radiol 2004; 11:876–880.
5. Skaane P, Young K, Skjennald A. Population-based mammography screening: comparison of screen-film and full-field digital mammography with soft-copy reading—Oslo I Study. Radiology 2003; 229: 877–884.
6. Skaane P, Skjennald A. Screen-film mammography versus full-field digital mammography with soft-copy reading: randomized trial in a population-based screening program—The Oslo II Study. Radiology 2004; 232:197–204.
7. Skaane P, Balleyguier C, Diekmann F, et al. Breast lesion detection and classification: comparison of screen-film mammography and full-field digital mammography with soft-copy reading—observer performance study. Radiology 2005; 237:37–44.
8. Hillman BJ, Fajardo LL, Hunter TB, et al. Mammogram interpretation by physician assistants. AJR Am J Roentgenol 1987; 149:907–911.
9. Bassett LW, Bunnell DH, Jahanshahi R, et al. Breast cancer detection: one versus two views. Radiology 1987; 165:95–97.
10. Wallis MG, Walsh MT, Lee JR. A review of false negative mammography in a symptomatic population. Clin Radiol 1991; 44:13–15.
11. Harvey JA, Fajardo LL, Innis CA. Previous mammograms in patients with impalpable breast carcinomas: retrospective vs blinded interpretation. AJR Am J Roentgenol 1993; 161:1167–1172.
12. Bird RE, Wallace TW, Yankaskas BC. Analysis of cancers missed at screening mammography. Radiology 1992; 184:613–617.
13. Beam CA, Layde PM, Sullivan DC. Variability in the interpretation of screening mammograms by US radiologists—findings from a national sample. Arch Intern Med 1996; 156:209–213.
14. Beam V, Sullivan D, Layde P. Effect of human variability on independent double reading in screening mammography. Acad Radiol 1996; 3:891–897.
15. Shtern F, Stelling C, Goldberg B, et al. Novel technologies in breast imaging: National Cancer Institute perspective. In: Society of Breast Imaging Conference. Orlando, FL, 1995; 153–156.
16. Vyborny CJ. Can computers help radiologists read mammograms? Radiology 1994; 191:315–317.
17. Li L, Clark RA, Thomas JA. Computer-aided diagnosis of masses with full-field digital mammography. Acad Radiol 2002; 9:4–12.
18. Petrick N, Chan HP, Sahiner B, et al. Breast cancer detection: evaluation of a mass detection algorithm for computer-aided diagnosis: experience in 263 patients. Radiology 2002; 224:217–224.
19. Wei J, Sahiner B, Hadjiiski LM, et al. Two-view information fusion for improvement of computer-aided detection (CAD) of breast masses on mammograms. SPIE Proc 2006; 6144:241–247.
20. Wei J, Sahiner B, Chan HP, et al. Computer aided diagnosis system for mass detection: comparison of performance on full-field digital mammograms and digitized film mammograms. RSNA 2003, Chicago, November 30–December 5, 2003:387.
21. Wei J, Sahiner B, Hadjiiski LM, et al. Computer aided detection of breast masses on full field digital mammograms. Med Phys 2005; 32: 2827–2838.
22. Burt PJ, Adelson EH. The Laplacian pyramid as a compact image code. IEEE Trans Commun 1983; COM-31:337–345.
23. Wei J, Sahiner B, Hadjiiski LM, et al. Computer aided detection of breast masses on full-field digital mammograms: false positive reduction using gradient field analysis. Proc SPIE Med Imaging 2004; 5370: 992–998.
24. Petrick N, Chan HP, Sahiner B, et al. Combined adaptive enhancement and region-growing segmentation of breast masses on digitized mammograms. Med Phys 1999; 26:1642–1654.
25. Sahiner B, Petrick N, Chan HP, et al. Computer-aided characterization of mammographic masses: accuracy of mass segmentation and its effects on characterization. IEEE Tran Med Imaging 2001; 20:1275–1284.
26. Wei D, Chan HP, Petrick N, et al. False-positive reduction technique for detection of masses on digital mammograms: global and local multiresolution texture analysis. Med Phys 1997; 24:903–914.
27. Norusis MJ. SPSS for Windows release 6 professional statistics. Chicago, IL: SPSS Inc, 1993.
28. Hadjiiski LM, Sahiner B, Chan HP, et al. Analysis of temporal change of mammographic features: computer-aided classification of malignant and benign breast masses. Med Phys 2001; 28:2309–2317.
29. Metz CE, Herman BA, Shen JH. Maximum-likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data. Stat Med 1998; 17:1033–1053.
30. Sahiner B, Chan HP, Petrick N, et al. Feature selection and classifier performance in computer-aided diagnosis: the effect of finite sample size. Med Phys 2000; 27:1509–1522.
31. Swensson RG. Unified measurement of observer performance in detection and localizing target objects on images. Med Phys 1996; 23: 1709–1724.
32. Chakraborty DP, Winter LHL. Free-response methodology: alternate analysis and a new observer-performance experiment. Radiology 1990; 174:873–881.
33. Chakraborty DP, Berbaum KS. Observer studies involving detection and localization: modeling, analysis, and validation. Med Phys 2004; 31:2313–2330.
34. Hendrick R, Lewin J, D'Orsi C, et al. Non-inferiority study of FFDM in an enriched diagnostic cohort: comparison with screen-film mammography in 625 women. In: Yaffe MJ, ed. IWDM 2000: 5th International Workshop on Digital Mammography: Medical Physics, 2001; 475–481.
35. Chan HP, Wei J, Zhou C, Helvie MA, et al. Comparison of mammographic density estimated on digital mammograms and screen-film mammograms. RSNA 89th Scientific Assembly (Chicago, IL) 2003:424.
36. Chan HP, Sahiner B, Wagner RF, et al. Classifier design for computer-aided diagnosis: effects of finite sample size on the mean performance of classical and neural network classifiers. Med Phys 1999; 26:2654–2668.
37. Kupinski MA, Giger ML. Feature selection with limited datasets. Med Phys 1999; 26:2176–2182.
38. O'Shaughnessy KF, Castellino RA, Muller SL, et al. Computer-aided detection (CAD) on 90 biopsy-proven breast cancer cases acquired on a full-field digital mammography (FFDM) system. Radiology 2001; 221: 471.
39. Warren Burhenne LJ, Wood SA, D'Orsi CJ, et al. Potential contribution of computer-aided detection to the sensitivity of screening mammography. Radiology 2000; 215:554–562.
40. Brem RE, Hoffmeister JW, Rapelyea JA, et al. Impact of breast density on computer-aided detection for breast cancer. AJR Am J Roentgenol 2005; 184:439–444.

Bilateral analysis based false positive reduction for computer-aided mass detection

Yi-Ta Wu

Jun Wei

Lubomir M. Hadjiiski

Berkman Sahiner

Chuan Zhou

Jun Ge

Jiazheng Shi

Yiheng Zhang

Heang-Ping Chan

Department of Radiology

University of Michigan, Ann Arbor, MI 48109

Correspondence:

Yi-Ta Wu, Ph.D.
Department of Radiology
University of Michigan
CGC B2103
1500 E. Medical Center Drive
Ann Arbor, MI 48109-0904

Phone: 734-647-8553
Fax: 734-615-5513
Email: yitawu@umich.edu

Running head: Bilateral mammogram based CAD system for masses

ABSTRACT

We have developed a false positive (FP) reduction method based on analysis of bilateral mammograms for computerized mass detection systems. The mass candidates on each view were first detected by our unilateral computer-aided detection (CAD) system. For each detected object, a regional registration technique was used to define a region of interest (ROI) that is "symmetrical" to the object location on the contralateral mammogram. Texture features derived from the spatial gray level dependence (SGLD) matrices and morphological features were extracted from the ROI containing the detected object on a mammogram and its corresponding ROI on the contralateral mammogram. Bilateral features were then generated from corresponding pairs of unilateral features for each object. Two linear discriminant analysis (LDA) classifiers were trained from the unilateral and the bilateral feature spaces, respectively. Finally, the scores from the unilateral LDA classifier and the bilateral LDA asymmetry classifier were fused with a third LDA whose output score was used to distinguish true mass from false positives (FPs). A data set of 341 cases of bilateral two-view mammograms was used in this study, of which 276 cases with 552 bilateral pairs contained 110 malignant and 166 benign biopsy-proven masses and 65 cases with 130 bilateral pairs were normal. The mass data set was divided into two subsets for 2-fold cross-validation training and testing. The normal data set was used for estimation of FP rates. It was found that our bilateral CAD system achieved a case-based sensitivity of 70%, 80%, and 85% at average FP rates of 0.35, 0.75, and 0.95 FPs/image, respectively, on the test data sets with malignant masses. In comparison to the average FP rates for the unilateral CAD system of 0.58, 1.33, and 1.63, respectively, at the corresponding sensitivities, the FP rates were reduced by 40%, 44%, and 42% with the bilateral symmetry information. The improvement was statistically significance ($p<0.05$) as estimated by JAFROC analysis.

Keywords: computer-aided detection (CAD), bilateral analysis, mass detection, false positive reduction.

I. INTRODUCTION

Breast cancer is one of the leading causes of death among American women between 40 to 55 years of age.[1] It has been reported that early diagnosis and treatment can improve significantly the chance of survival for patients with breast cancer.[2-4] Although mammography is a powerful screening tool for detecting breast cancer,[5, 6] studies indicate that a substantial fraction of breast cancers that are visible upon retrospective analyses of the images are not detected initially.[7-9] It has been shown that computer-aided detection (CAD) can increase the cancer detection rate by radiologists both in the laboratory and in clinical practice.[10-15]

In screening mammography, two mammographic views, cranio-caudal (CC) and mediolateral oblique (MLO) views are generally taken of each breast. During mammographic interpretation, the radiologist combines complex information including morphology, texture, and geometric location of any suspicious structures of the imaged breast from different views, asymmetric density patterns between bilateral mammograms of the same view, and changes between the current and the prior mammograms if available. Radiologists have found that these techniques are effective in improving the accuracy of detecting subtle lesions and reducing false positives (FPs).

Investigators have attempted to implement the multiple image techniques in CAD systems to improve the detection accuracy of abnormalities and the classification accuracy of differentiating malignant and benign lesions. Hadjiiski *et al.*[16] developed an interval change analysis of masses on current and prior mammograms and found that the classification accuracy of masses can be improved significantly in comparison to single image classification. Paquerault *et al.*[17] developed a two-view (CC and MLO views) fusion technique to reduce FPs in mass detection and obtained significant improvement by comparing to their one-view detection system. Engeland *et al.*[18] recently presented a two-view CAD system by using the features including the difference in the radial

distance from the candidate regions to the nipple, the gray scale correlation between both regions, and the mass likelihood of the regions determined by the single view CAD scheme. Yin *et al*.[19] used bilateral subtraction in a prescreening step of a mass detection program to locate mass candidates, but the subsequent image analysis was performed based only on a single view. Mendez *et al*.[20] developed a bilateral CAD system based on a bilateral subtraction approach and used size and eccentricity tests and texture features to eliminate FPs. Again, the bilateral information is only used to find the suspicious objects and the subsequent analysis is based on a single view.

The detection of masses on mammograms is a challenging task. The normal fibroglandular tissue in the breast causes FPs by mimicking masses and causes false negatives (FNs) due to overlapping with lesions. In order to improve the performance of our mass detection system, we are investigating computer-vision methods by incorporating information from two-view mammograms[17] and bilateral mammograms,[21] emulating radiologists' mammographic interpretation techniques. In this study, we will discuss our approach to FP reduction by analyzing the symmetry or asymmetry of density patterns between bilateral mammograms.

II. MATERIALS AND METHODS

A. Data Sets

A database of mammograms was collected from patient files at the Department of Radiology with Institutional Review Board (IRB) approval. The mammograms were digitized by a Lumiscan laser scanner with a pixel size of $50\mu m \times 50\mu m$ and 12 bits per pixel. The pixel size was increased to $100\mu m \times 100\mu m$ by averaging every 2×2 adjacent pixels before being input to the CAD system. In this study, two data sets are used: a mass data set containing bilateral digitized mammograms with malignant or benign masses and a no-mass data set containing bilateral digitized mammograms

without masses, verified by an experienced radiologist. All cases had four mammographic views, the CC view and the MLO view mammogram for both breasts. The mass set and the no-mass data set contained 276 cases (552 bilateral pairs) and 65 cases (130 bilateral pairs), respectively, yielding a total of 1364 mammograms. The mass data set was used to estimate the detection sensitivity and the no-mass data set was used for estimating the FP rate (number of FPs per image). In the mass data set, each patient had a biopsy-proven mass in one of the breasts, resulting in a total of 276 masses, 166 of which were benign and 110 malignant. An MQSA radiologist identified the location of the masses based on all available diagnostic and clinical information of the case, measured the mass sizes as the longest dimension seen on the two-view mammograms, provided descriptors of the mass shapes and mass margins, and also provided an estimate of the breast density in term of BI-RADS category. Figure 1 shows the information of our data set which includes the distributions of mass sizes, mass shapes, mass margins, and breast density.

For training and evaluation of the performances of the CAD systems, the cases in our mass data set were divided into two independent data subsets containing 136 and 140 cases, respectively, for two-fold cross validation training and testing. Of the 136 cases in subset 1, 52 were malignant and 84 were benign. Of the 140 cases in subset 2, 58 were malignant and 82 were benign. The no-mass data set was not used during training. All 260 mammograms were kept as independent test samples to be used with both test subsets.

B. Methods

Our bilateral CAD system combines unilateral features with bilateral features to reduce FPs. Similar structures that appear in both right and left mammograms at corresponding locations are more likely to be normal tissue than masses, whereas asymmetric density may indicate a developing lesion. The

key of this system is therefore the design of a classifier that can differentiate symmetry and asymmetry of paired ROIs in corresponding regions on bilateral mammograms of the same view. The system consists of four steps: (1) mass candidate (MC) localization, (2) corresponding ROIs (CR) registration, (3) feature extraction and analysis, and (4) bilateral information fusion. Figure 2 shows the block diagram for our bilateral CAD system. The detailed description for each step is presented below.

1. Mass Candidate Localization

Identification of mass candidates is performed by the following two steps: breast segmentation and mass candidate detection. The breast image is first segmented from the surrounding image background by boundary detection.

The algorithm developed by Zhou *et al.*[22] in our laboratory is used to track the breast boundary and segment the breast from the background. Mass detection is performed only in the breast region. We have previously developed a mass detection system for unilateral mammograms.[23-25] The system is used for mass candidate detection in the current study. The system performs mass detection in two steps. In the first step, a gradient field analysis method is used to determine the seeds of mass candidates followed by a region growing[24] method to segment the mass candidates starting from those seeds. In the second step, the gradient convergence is calculated using the gray levels and the shape of the segmented mass region as a priori information. The mass candidates that pass the gradient convergence criterion are retained for further analysis in the bilateral system. Figure 3 shows an example of mass candidates detected on a mammogram. Figure 3(a), 3(b), and 3(c) show the original image, detected breast boundary, and the detected mass candidates, respectively.

2. Corresponding ROI Registration

For each mass candidate, its corresponding ROI on the contralateral mammogram is identified by the regional registration technique developed previously in our laboratory[16] with a modification to handle the special case when the distance between the nipple location and the center of an ROI is too small to obtain the intersection points on the breast boundary. The nipple location on each image was manually identified so that the effectiveness of the bilateral analysis method could be evaluated independent of nipple detection errors.

The original region registration technique included the following steps. The registration is performed in a polar coordinate system where the origin is located at the nipple location of a breast image. Figure 4 shows an example of locating the corresponding ROI of a mass candidate on the contralateral mammogram. Using the distance *r* from the nipple *o* to the center of the mass as the radius, an arc centered at the origin (nipple) is drawn. The arc will intersect the mass candidate and the breast boundary at two points, *p* and *q*. The angle between $\overrightarrow{om}$ and $\overrightarrow{op}$ is defined as θ, the angle between $\overrightarrow{op}$ and $\overrightarrow{oq}$ is defined as α. On the contralateral mammogram, the corresponding ROI *m'* is localized with a similar procedure. An arc of radius *r* centered at the nipple *o'* of the contralateral mammogram is drawn. The intersections of the arc with the breast boundary are *p'* and *q'*. The angle between $\overrightarrow{o'p'}$ and $\overrightarrow{o'q'}$ is defined as α'. The location of the corresponding ROI as determined by the angle θ' between $\overrightarrow{o'p'}$ and the radius $\overrightarrow{o'm'}$ is estimated as $\theta\alpha'/\alpha$. The coordinate of the center of the corresponding ROI is therefore given by (r, θ').

For some special cases that the nipple is located within the breast, not on the breast boundary (referred to as an inward nipple), our original regional registration method may fail since the

distance from the nipple to the mass candidate can be too short to obtain two intersection points on the breast boundary. In order to handle those special cases, the new origin will be derived by horizontally shifting the origin of the polar coordinate system toward the breast boundary until the intersection on the breast boundary is reached. In this way, the radius can be roughly determined such that the corresponding ROI location can be estimated. Figure 5 shows an example of the modified regional registration. Figure 5(a) shows an example that the distance $\overrightarrow{om}$ between the nipple and a mass candidate is too small to obtain two intersection points at the breast boundary. After horizontally shifting the origin from *o* to *n* in Figure 5(a) and the origin from *o'* to *n'* in Figure 5(b), the location of the corresponding ROI *m'* is estimated based on the new origins using the regional registration technique as described above.

3. Feature Extraction and Analysis

3.1 Feature Extraction

For the features analysis, two types of features, SGLD (spatial gray-level dependence) texture features and morphological features are extracted from both the ROI containing the detected mass candidate and its contralateral ROI.

For the SGLD features, thirteen texture measures[24-26] are extracted from the entire ROI (referred to as the global texture features) at 14 distances and 2 angles with a total of 364 (13x14x2) features. The same 13 texture measures are extracted from the central region containing the detected object and the peripheral regions within each ROI (referred to as the local texture features) at 4 distances and 2 angles with a total of 104 (13x2x4) features from the central region and 104 features as the difference of the corresponding features in the central and the peripheral regions.[25]

Twelve morphological features are extracted from the object segmented within the ROI.[24, 25] Five of them are based on the normalized radial length (NRL), defined as the Euclidean distance from the object centroid to each of its edge pixels and normalized relative to the maximum radial length for the object.[27] In our previous studies, we found that the mean, standard deviation, entropy, area ratio, and zero crossing count features derived from the NRL are useful for discriminating between objects containing masses and normal tissue.[24] The other six morphological features are the perimeter, area, perimeter-to-area ratio, circularity, rectangularity, and contrast of the object.[24] The last morphological feature is the summary Fourier descriptor measure,[28] which is obtained from the Fourier transform of the object boundary sequence. Objects with more irregular contours have more high-frequency components than those with smooth contours.[29]

3.2 Unilateral CAD System

The unilateral LDA classifier uses only the SGLD texture features as input predictor variables as described previously.[25] The stepwise LDA feature selection strategy with simplex optimization[16] was used to select the best texture feature subset and reduce the dimensionality of the feature space. Two-fold cross validation was used to train and test the CAD systems, as discussed below. For each of the two cross validation cycles, the algorithm used a leave-one-case-out resampling method and simplex optimization within the training subset to estimate the best threshold values, F_{in}, F_{out}, and tolerance, based on the F statistics for stepwise feature selection. The chosen F_{in}, F_{out} and tolerance values are then used to select a set of features and the weights for the LDA classifier are estimated from the training subset. The test subset was thus independent of the classifier training in each cross-validation cycle. This procedure has been described in more details previously.[23]

4. Bilateral Information Fusion

4.1 Bilateral LDA Classifier

The bilateral LDA classifier incorporates the "symmetry" information on the left and right breasts to differentiate symmetric (likely FPs) and asymmetric (likely masses) structures. Bilateral features are derived from the unilateral SGLD texture features and the morphological features for each pair of ROIs – a detected mass candidate and its corresponding ROI, using the following relationship:

$$BF[i, j] = \frac{Max(MC[i, j], CR[i, j])}{Min(MC[i, j], CR[i, j])} \quad (1)$$

where $MC[i, j]$ and $CR[i, j]$ are the i^{th} feature of the j^{th} mass candidate and the i^{th} feature of the j^{th} corresponding ROI, respectively. The bilateral LDA classifier was trained in a similar way as that for the unilateral LDA classifier, as described above.

4.2 Bilateral CAD System

In the last stage, the discriminant scores of the unilateral and bilateral classifiers are merged by a third LDA. The weights of this LDA classifier were also trained with the training subset. The output score from the third LDA is used to differentiate TPs from FPs in the bilateral CAD system.

5. Evaluation Methods

The detected individual objects were compared with the true mass location marked by an experienced radiologist. An object was considered to be a true positive (TP) if the overlap between the detected object and the true mass was greater than 25%. The 25% threshold was selected as described in our previous study.[30]

To evaluate the performance of our bilateral LDA classifier, the test discriminant scores were analyzed using receiver operating characteristic (ROC) methodology.[31] The accuracy for classification of mass and normal tissue was evaluated as the area under the ROC curve, A_z.

The detection performance of the bilateral CAD system was assessed by free response ROC (FROC) analysis. An FROC curve shows the relationship between the detection sensitivity and the FP rate as the decision threshold varies. FROC curves were presented on a per-image and a per-case basis. For image-based FROC analysis, the mass on each mammogram was considered an independent true object. For case-based FROC analysis, the same mass imaged on the two-view mammograms was considered to be one true object and detection of the masses on either view or on both views was considered to be a TP detection.

Two sets of trained parameters were acquired as a result of the 2-fold cross validation training. To estimate the FP rate on normal mammograms when the trained CAD system is used in a screening setting, we applied the trained unilateral and bilateral systems to the 260 no-mass mammograms for independent testing. The number of FP marks produced by the algorithm was estimated by counting the detected objects on these normal cases only. The mass sensitivity was determined by counting only the masses on the corresponding test mass subset. The combination of the sensitivity from the test mass subset and the FP rate from the normal data set at the corresponding detection thresholds resulted in a test FROC curve. The training and testing procedure were performed for each cycle of the two-fold cross validation process, thereby generating two test FROC curves. To estimate the overall performance of the CAD system, an average test FROC curve is obtained by averaging the FP rates from the FROC curves of the two mass subsets at the corresponding sensitivities.

Chakraborty *et al.*[32] proposed a JAFROC method and provided software to estimate the statistical significance of the difference between two FROC curves. We employed the JAFROC analysis to evaluate the difference in the FROC curves obtained from the unilateral CAD system and the bilateral CAD system.

III. RESULTS

A. Bilateral Feature Analysis

Figures 6 and 7 show examples of detection results obtained from the unilateral system and the bilateral system. Figure 6 shows a mass that was initially detected as a mass candidate but was excluded in the false positive reduction steps, and was therefore an FN of the unilateral CAD system. The bilateral analysis increased the likelihood score of this mass. It was therefore not excluded in the false positive reduction steps and became a TP in the bilateral CAD system.

Figure 7 shows an example of an FP detected by the unilateral CAD system. The FP was excluded in the bilateral system because it was found to have high symmetry with the tissue in the contralateral breast, as shown in the ROI in Figure 7(d), by the bilateral analysis.

B. Performance Evaluation

In the prescreening process, we obtained a large number of mass candidates on each mammogram. Each mass candidate was paired with a corresponding ROI in the contralateral breast. A total of 3127 and 3402 mass candidates were extracted for training subset 1 and subset 2, respectively, which included 98.5% (134/136) and 99.3% (139/140) of the masses in the two subsets. The mass candidates in the unilateral mammograms and the ROI pairs from bilateral mammograms in the training subset were used to design the unilateral and bilateral classifiers in each of the 2-fold cross-

validation cycles. The most effective subset of features from the available feature pool was selected for each of the training subsets during the training procedure. For the unilateral LDA classifier, twenty (11 global and 9 local) and nineteen (12 global and 7 local) texture features were selected from the two independent training subsets, respectively. For the bilateral LDA classifier, twenty-four (11 global texture, 9 local texture and 4 morphological) and twenty-three (12 global, 8 local, and 3 morphological) features were selected from the two independent training subsets, respectively. The validation A_z values of the LDA classifier during the leave-one-case-out training were 0.846 ± 0.011 and 0.832 ± 0.009, respectively, for the two training subsets using the unilateral LDA classifier, and were 0.862 ± 0.015 and 0.859 ± 0.012, respectively, using the bilateral LDA classifier. The classifiers achieved A_z values of 0.833 ± 0.015 and 0.831 ± 0.011, respectively, for the two test subsets using the unilateral LDA classifier, and 0.853 ± 0.013 and 0.849 ± 0.011, respectively, using the bilateral LDA classifier.

Figure 8 shows the average test FROC curves for the unilateral and bilateral CAD systems after FP reduction with the corresponding trained LDA classifiers when the FP rates were estimated from the test subsets with masses. Figure 9 shows the corresponding results when the FP rates were estimated on the set of no-mass mammograms. Table I summarizes the average FP rates estimated with both the mass and no-mass data sets at several case-based sensitivities.

Because the detection performance of CAD systems on cancer cases is of prime importance, we analyzed the performance of our CAD systems for the subset of cases containing malignant masses. Figure 10 compares the average test FROC curves for the unilateral and bilateral CAD systems on malignant cases only. Figure 11 shows the average test FROC curves for the unilateral and bilateral CAD systems with the sensitivities estimated on malignant cases only and the FP rates estimated on the set of no-mass mammograms. The bilateral CAD system achieved a case-based

sensitivity of 70%, 80%, and 85% at average FP rates of 0.35, 0.75, and 0.95 FPs/image, respectively, on the test subset of malignant masses. In comparison to the average FP rates for the unilateral CAD system of 0.58, 1.33, and 1.63 FPs/image, respectively, at the corresponding sensitivities, the FP rates were reduced by 40%, 44%, and 42% with the bilateral symmetry information. Table II summarizes the average FP rates estimated with both the mass and no-mass data sets for cases with malignant masses only at several case-based sensitivities.

The figure-of-merit (FOM) from the output of the JAFROC software is summarized in Table III(a) for all cases and in Table III(b) for malignant cases only. The difference between the FOMs for the unilateral and the bilateral CAD systems was statistically significant ($p<0.05$) for all comparisons.

IV. DISCUSSION

Symmetry between breast structures in bilateral pairs of mammograms is an important feature used by radiologists for mass detection or FP reduction. Similar structures that appear in both right and left mammograms are more likely to be normal tissue than abnormal lesions. Our bilateral analysis translates this radiologists' knowledge to computer vision techniques so that the CAD system can utilize the symmetry of breast tissue on bilateral mammograms to improve detection accuracy. The results of our study show that the bilateral information is an effective technique for reducing FPs.

The bilateral features are important factors affecting the performance of the bilateral LDA classifier. In this study, the bilateral features were derived from features extracted from each pair of ROIs, i.e., the mass candidate and its corresponding ROI, using the maximum-to-minimum ratio strategy as shown in Eq. (1). We also investigated if other strategies, including $BF[i,j]=\frac{MC[i,j]}{CR[i,j]}$,

$BF[i,j]=\frac{MC[i,j]-CR[i,j]}{MC[i,j]}$, and $BF[i,j]=\frac{MC[i,j]-CR[i,j]}{(MC[i,j]+CR[i,j])/2}$, could improve the performance of the bilateral CAD system. It was found that these strategies are not as effective as the maximum-to-minimum ratio. Specifically, among the A_z values of all bilateral features, 72% of those from the latter strategies are lower than those of their corresponding features obtained by Eq. (1). The advantage of using bilateral symmetry measures defined by the maximum-to-minimum ratio can be seen by considering the following example: assuming two ROI pairs that are highly asymmetric, (MC_1, CR_1) and (MC_2, CR_2), in which $MC_1 > CR_1$ and $MC_2 < CR_2$, their bilateral features derived as the maximum-to-minimum ratio will both be greater than 1. However, the bilateral features obtained from $BF[i,j]=\frac{MC[i,j]}{CR[i,j]}$ will be greater than 1 for (MC_1, CR_1) but smaller than 1 for (MC_2, CR_2). The bilateral measures obtained from $BF[i,j]=\frac{MC[i,j]-CR[i,j]}{MC[i,j]}$ or $BF[i,j]=\frac{MC[i,j]-CR[i,j]}{(MC[i,j]+CR[i,j])/2}$ will be positive for (MC_1, CR_1) but negative for (MC_2, CR_2). The bilateral feature defined in Eq. (1) therefore describes the asymmetry between the ROI pairs, regardless which ROI has a larger feature value, whereas the other three bilateral features do not consistently provide feature values in the same direction. The maximum-to-minimum ratio approach can thus achieve better performance than the other three strategies.

The corresponding ROI registration is an important procedure in the bilateral analysis. The two breasts of a given patient are not perfectly symmetrical and other factors such as positioning and compression further introduce variability in the symmetry. We investigated the effect of variability in the registered ROI locations on bilateral analysis. For this purpose, the pre-screening step of our unilateral CAD system was first applied to the contralateral mammogram to locate the mass candidates. For a given ROI predicted by the registration method on the contralateral mammogram, its location was compared to the ROI locations of these mass candidates by evaluating an overlap

ratio, defined as the intersection between the predicted ROI and a mass candidate ROI relative to the area of the smaller ROIs. If the overlap ratio of the predicted ROI with a mass candidate ROI was greater than a chosen threshold, the location of the predicted ROI would be changed to the location of the mass candidate ROI. If the predicted ROI overlapped with more than one mass candidate ROIs, the mass candidate ROI having the largest overlap ratio that exceeded the threshold would be used. We evaluated the effects of this ROI location adjustment for a range of thresholds. It was found that when the overlap ratio threshold was chosen to be about 0.7 to 0.9, the performance of the bilateral CAD system would have a small but insignificant improvement compared to the bilateral CAD system without the ROI adjustment process. When the overlap ratio threshold was smaller than 0.5, the performance of the bilateral CAD system was degraded. This study indicated that small variability of the predicted ROI location on the contralateral mammogram does not have a strong effect on the performance of the bilateral analysis.

Various registration methods have been attempted for registration of mammograms of the same breast. For example, the warping approach proposed by Sallam et al.[33], and the multiple-control-point approach proposed by Vujovic et al.[34]. Those approaches depended on the identification of corresponding control points. However, there are few, if any, invariant landmarks on mammograms that can be identified automatically because the breast is composed of soft tissue. The projected image of the breast tissue often changes even when the same breast is compressed two different times. It is even more variable between a breast and its contralateral breast. Commonly used rigid or non-rigid registration methods will not be appropriate for this application. We therefore developed the regional registration method for correlation of ROIs on mammograms. Our regional registration method uses the nipple and the distance between the nipple and the ROI center to be the relatively invariant information. The lesion in the target breast is estimated to be located

within a band of tissue centered along the arc traced using the nipple-to-lesion distance as the radius and with the origin at the nipple. This method emulates a technique used by many radiologists in identifying corresponding lesions in two-view mammograms or current and prior mammograms. Van Engeland et al.[35] compared methods for mammogram registration based on breast alignment and linear and non-linear warping. They concluded that linear warping using mutual information performed better than the other methods. We also performed a study comparing our regional registration method to correlation or mutual information based linear and non-linear warping methods using a data set of 390 current and prior mammogram pairs[36]. Our results showed that the regional registration method outperformed the warping approaches in identifying corresponding lesions on the mammogram pairs. The localization of symmetric ROIs on the bilateral breasts is similar to the problem of registering ROIs on current and prior mammograms. We therefore adapted the regional registration method to the bilateral analysis in this study.

To implement the bilateral analysis in a practical CAD system, the nipple locations have to be detected automatically. We have previously developed a nipple detection algorithm to determine the nipple location on a mammogram. The algorithm could detect the nipple locations within 1 cm of the manually identified locations in about 70% of the images in the data set used in this study. A large deviation of the nipple location from the true location may affect the regional registration technique in locating the symmetric ROI on the contralateral mammogram, which in turn may degrade the performance of the bilateral analysis of tissue symmetry. We therefore used the manually identified nipple locations in this study in order to develop the bilateral classifier without the influence of other confounding factors. Further work is underway to improve the nipple detection algorithm and to investigate the effect of nipple detection accuracy on the performance of the bilateral system.

The inward nipple projection is often a result of positioning and compression problems so that the nipple is not projected in profile. Since there is not enough information from the 2D projected mammograms to correct for the deformation of the breast, we designed a simple, *ad hoc* correction method to allow the arc drawn using the nipple-to-mass distance as the radius to intersect the breast boundary. In these cases, the breast image on the bilateral mammogram often does not have a similar positioning problem and the difference in the compression of the two breasts may cause large uncertainty in the registration regardless of the correction method. For cases in which both breasts actually have inward nipples and the breast images are similar, our correction method will not cause additional errors because similar correction will be applied to the bilateral mammograms and symmetric ROIs will be identified on the mammograms.

Our motivation of this study is to reduce the FPs of a CAD system for mass detection. A CAD detection system is generally intended for use in screening mammography. At the screening stage, all lesions of concern should be pointed out to radiologists so that the radiologists can judge whether a recall is warranted. If a detection system is trained to mark only the malignant lesions, it may be attempting to play the role of a triage system (alerting radiologists to work up only "malignant" cases) rather than that of a second reader. Furthermore, since computerized lesion detection or characterization on mammograms is not 100% sensitive, it will be confusing to the radiologists whether an unmarked suspicious lesion is missed or it is considered benign by the computer. We believe that computer-aided diagnosis (CADx) may be used in different ways in conjunction with a CAD detection system. For example, the likelihood of malignancy may be estimated by the CADx system and displayed for every detected lesion, and/or a CADx system may be used during diagnostic workup. Either way the CAD system will first alert radiologists to all masses, leaving the assessment of malignancy or benignity to a second stage. We therefore

included both malignant and benign masses in the training sets to train the system to detect all masses.

V. CONCLUSIONS

We developed an FP reduction method to improve computerized mass detection on mammograms based on analysis of bilateral information. It was found that the false positives can be reduced by training a new classifier for bilateral features and combining its output score with the unilateral classifier score. The bilateral CAD system achieved a case-based sensitivity of 70%, 80%, and 85% for detection of malignant masses at average FP rates of 0.35, 0.75, and 0.95 FPs/image, respectively, on the test data set. In comparison to the average FP rates for the unilateral CAD system of 0.58, 1.33, and 1.63 FPs/image, respectively, at the corresponding sensitivities, the FP rates were reduced by 40%, 44%, and 42% with the bilateral symmetry information. The improvement in the overall detection accuracy is statistically significant ($p<0.05$) by JAFROC analysis. Our results demonstrate that the bilateral analysis can differentiate the similarity and dissimilarity between tissues at corresponding locations in the bilateral views, and is useful for improving the performance of a unilateral CAD system by further reducing the FPs.

ACKNOWLEGMENTS

This work is supported by USPHS grant CA95153 and U. S. Army Medical Research and Materiel Command grants DAMD17-02-1-0214 and W81XWH-1-04-1-0475. The content of this paper does not necessarily reflect the position of the government and no official endorsement of any equipment and product of any companies mentioned should be inferred. The authors are grateful to

Charles E. Metz, Ph.D., for the LABROC program, and to Dev Chakraborty, Ph.D., for the JAFROC program.

REFERENCES

[1]"American Cancer Society," www.cancer.org 2004, Statistics for 2004.

[2]C. R. Smart, R. E. Hendrick, J. H. Rutledge, and R. A. Smith, "Benefit of mammography screening in women ages 40 to 49 years: current evidence from randomized controlled trials," Cancer **75**, 1619-1626 (1995).

[3]S. A. Feig, C. J. D'Orsi, R. E. Hendrick, V. P. Jackson, D. B. Kopans, B. Monsees, E. A. Sickles, C. B. Stelling, M. Zinninger, and P. Wilcox-Buchalla, "American College of Radiology guidelines for breast cancer screening," AJR Am J Roentgenol. **171**, 29-33 (1998).

[4]B. Cady and J. S. Michaelson, "The life-sparing potential of mammographic screening," CANCER **91**, 1699-1703 (2001).

[5]L. Tabar, C. Fagerberg, A. Gad, L. Baldetorp, L. Holmberg, O. Grontoft, U. Ljungquist, B. Lundstrom, J. Manson, G. Eklund, and e. al, "Reduction in mortality from breast cancer after mass screening with mammography," Lancet **1**, 829-832 (1985).

[6]H. C. Zuckerman, The role of mammography in the diagnosis of breast cancer. *In: Breast Cancer, Diagnosis and Treatment*, by I. M. Ariel and J. B. Cleary, (McGraw-Hill, New York, 1987).

[7]C. A. Beam, P. M. Layde, and D. C. Sullivan, "Variability in the interpretation of screening mammograms by US radiologists - Findings from a national sample," Archives of Internal Medicine **156**, 209-213 (1996).

[8]R. L. Birdwell, D. M. Ikeda, K. F. O'Shaughnessy, and E. A. Sickles, "Mammographic characteristics of 115 missed cancers later detected with screening mammography and the potential utility of computer-aided detection," Radiology **219**, 192-202 (2001).

[9]J. G. Elmore, C. Y. Nakano, T. D. Koepsell, L. M. Desnick, C. J. D'Orsi, and D. F. Ransohoff, "International variation in screening mammography interpretations in community-based programs," J. National Cancer Institute **95**, 1384-1393 (2003).

[10]L. J. Warren Burhenne, S. A. Wood, C. J. D'Orsi, S. A. Feig, D. B. Kopans, K. F. O'Shaughnessy, E. A. Sickles, L. Tabar, C. J. Vyborny, and R. A. Castellino, "Potential contribution of computer-aided detection to the sensitivity of screening mammography," Radiology **215**, 554-562 (2000).

[11]T. W. Freer and M. J. Ulissey, "Screening mammography with computer-aided detection: Prospective study of 12,860 patients in a community breast center," Radiology **220**, 781-786 (2001).

[12]R. F. Brem, J. K. Baum, M. Lechner, S. Kaplan, S. Souders, L. G. Naul, and J. Hoffmeister, "Improvement in sensitivity of screening mammography with computer-aided detection: A multi-institutional trial," Am J Roentgenology **181**, 687-693 (2003).

[13]S. V. Destounis, P. DiNitto, W. Logan-Young, E. Bonaccio, M. L. Zuley, and K. M. Willison, "Can computer-aided detection with double reading of screening mammograms help decrease the false-negative rate? Initial experience," Radiology **232**, 578-584 (2004).

[14]H. P. Chan, K. Doi, C. J. Vyborny, R. A. Schmidt, C. E. Metz, K. L. Lam, T. Ogura, Y. Wu, and H. MacMahon, "Improvement in radiologists' detection of clustered microcalcifications on mammograms. The potential of computer-aided diagnosis," Investigative Radiology **25**, 1102-1110 (1990).

[15]M. A. Helvie, L. M. Hadjiiski, E. Makariou, H. P. Chan, N. Petrick, B. Sahiner, S. C. B. Lo, M. Freedman, D. Adler, J. Bailey, C. Blane, D. Hoff, K. Hunt, L. Joynt, K. Klein, C. Paramagul, S. Patterson, and M. A. Roubidoux, "Sensitivity of

noncommercial computer-aided detection system for mammographic breast cancer detection - A pilot clinical trial," Radiology **231**, 208-214 (2004).

[16]L. M. Hadjiiski, B. Sahiner, H. P. Chan, N. Petrick, M. A. Helvie, and M. N. Gurcan, "Analysis of Temporal Change of Mammographic Features: Computer-Aided Classification of Malignant and Benign Breast Masses," Medical Physics **28**, 2309-2317 (2001).

[17]S. Paquerault, N. Petrick, H. P. Chan, B. Sahiner, and M. A. Helvie, "Improvement of computerized mass detection on mammograms: Fusion of two-view information," Medical Physics **29**, 238-247 (2002).

[18]S. V. Engeland, S. Timp, and N. Karssemeijer, "Finding corresponding regions of interest in mediolateral oblique and craniocaudal mammographic views," Medical Physics **33**, 3203-3212 (2006).

[19]F. F. Yin, M. L. Giger, K. Doi, C. E. Metz, C. J. Vyborny, and R. A. Schmidt, "Computerized detection of masses in digital mammograms: Analysis of bilateral subtraction images," Medical Physics **18**, 955-963 (1991).

[20]A. J. Mendez, P. G. Tahoces, M. J. Lado, M. Souto, and J. J. Vidal, "Computer-aided diagnosis: Automatic detection of malignant masses in digitized mammograms," Medical Physics **25**, 957-964 (1998).

[21]Y.-T. Wu, L. M. Hadjiiski, J. Wei, C. Zhou, B. Sahiner, and H.-P. Chan, "Computer-aided detection of breast masses on mammograms: bilateral analysis for false positive reduction," Proc. SPIE, 2006, 6144, SPIE, 211-217.

[22]C. Zhou, H. P. Chan, N. Petrick, M. A. Helvie, M. M. Goodsitt, B. Sahiner, and L. M. Hadjiiski, "Computerized image analysis: Estimation of breast density on mammograms," Medical Physics **28**, 1056-1069 (2001).

[23]J. Wei, H.-P. Chan, B. Sahiner, L. M. Hadjiiski, M. A. Helvie, M. A. Roubidoux, C. Zhou, and J. Ge, "Dual system approach to computer-aided detection of breast masses on mammograms," Medical Physics **33**, 4157-4168 (2006).

[24]N. Petrick, H. P. Chan, B. Sahiner, and M. A. Helvie, "Combined adaptive enhancement and region-growing segmentation of breast masses on digitized mammograms," Medical Physics **26**, 1642-1654 (1999).

[25]J. Wei, B. Sahiner, L. M. Hadjiiski, H. P. Chan, N. Petrick, M. A. Helvie, M. A. Roubidoux, J. Ge, and C. Zhou, "Computer aided detection of breast masses on full field digital mammograms," Medical Physics **32**, 2827-2838 (2005).

[26]R. M. Haralick, K. Shanmugam, and I. Dinstein, "Texture features for image classification," IEEE Transactions on Systems, Man, and Cybernetics **SMC-3**, 610-621 (1973).

[27]J. Kilday, F. Palmieri, and M. D. Fox, "Classifying mammographic lesions using computer-aided image analysis," IEEE Transactions on Medical Imaging **12**, 664-669 (1993).

[28]L. Shen, R. M. Rangayyan, and J. E. L. Desautels, "Application of shape analysis to mammographic calcifications," IEEE Transactions on Medical Imaging **13**, 263-274 (1994).

[29]S. Mori, H. Nishida, and H. Yamada, *Optical Character Recognition*, (Wiley, New York, 1999).

[30]N. Petrick, H. P. Chan, B. Sahiner, M. A. Helvie, S. Paquerault, and L. M. Hadjiiski, "Breast cancer detection: Evalution of a mass detection algorithm for computer-aided diagnosis: Experience in 263 patients.," Radiology **224**, 217-224 (2002).

[31]C. E. Metz, B. A. Herman, and J. H. Shen, "Maximum-likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data," Statistics in Medicine **17**, 1033-1053 (1998).

[32]D. P. Chakraborty and K. S. Berbaum, "Observer studies involving detection and localization: modeling, analysis, and validation," Medical Physics **31**, 2313-2330 (2004).

[33]M. Sallam and K. Bowyer, Detecting abnormal densities in mammograms by comparison with previous screenings. *In: Digital Mammography '96*, by K. Doi, M. L. Giger, R. M. Nishikawa and R. A. Schmidt, (Elsevier, Amsterdam, 1996).

[34]N. Vujovic and D. Brzakovic, "Establishing the correspondence between control points in pairs of mammographic images," IEEE Trans Image Processing **6**, 1388-1399 (1997).

[35]S. Van Engeland, P. Snoeren, J. Hendriks, and N. Karssemeijer, "A comparison of methods for mammogram registration," IEEE Transactions on Medical Imaging **22**, 1436-1444 (2003).

[36]L. M. Hadjiiski, H. P. Chan, B. Sahiner, C. Zhou, M. A. Helvie, and M. A. Roubidoux, "Computerized Regional Registration of Corresponding Masses and Microcalcification Clusters on Temporal Pairs of Mammograms for Interval Change Analysis," Presented at the 89th Scientific Assembly and Annual Meeting of the Radiological Society of North America, Chicago, IL, November 30-December 5, 2003, RSNA Program Book 2003, RSNA, 389.

TABLE I. The average FP reduction rates at case-based sensitivities of 70%, 80%, and 85% for the test subsets when the FP rates were estimated from the mass and no-mass data sets.

	FP rate estimated from mass data set			**FP rate estimated from no-mass data set**		
	Unilateral CAD	**Bilateral CAD**	**FP Reduction**	**Unilateral CAD**	**Bilateral CAD**	**FP Reduction**
70%	0.70	0.53	24%	0.86	0.53	38%
80%	1.10	0.87	21%	1.32	1.04	21%
85%	1.46	1.15	21%	1.72	1.32	23%

570

TABLE II. The average FP reduction rates for cases with malignant masses at case-based sensitivities of 70%, 80%, and 85% for the test subsets when the FP rates were estimated from the mass and no-mass data sets.

	FP rate estimated from mass data set			**FP rate estimated from no-mass data set**		
	Unilateral CAD	**Bilateral CAD**	**FP Reduction**	**Unilateral CAD**	**Bilateral CAD**	**FP Reduction**
70%	0.43	0.33	23%	0.58	0.35	40%
80%	0.78	0.62	21%	1.33	0.75	44%
85%	0.94	0.78	17%	1.63	0.95	42%

TABLE III. Estimation of the statistical significance in the difference between the FROC performance of the unilateral and bilateral CAD systems on test subsets 1 and 2. The FP rates of the FROC curves were estimated from the no-mass data set: (a) all cases, and (b) malignant cases.

(a)

	FOM (JAFROC)	
	Test subset 1	**Test subset 2**
Unilateral CAD	0.52	0.48
Bilateral CAD	0.58	0.51
***p* value**	<0.001	0.008

(b)

	FOM (JAFROC)	
	Test subset 1 (malignant only)	**Test subset 2 (malignant only)**
Unilateral CAD	0.56	0.53
Bilateral CAD	0.61	0.56
***p* value**	0.009	0.003

575

FIGURE CAPTIONS

Figure 1. The characteristics of our mass data set. (a) distribution of mass sizes, (b) distribution of mass shapes, (c) distribution of mass margins, C: circumscribed, Ind: indistinct, M: microlobulated, Ob: obscured, Sp: spiculated, (d) distribution of the breast density in terms of BI-RADS category estimated by an MQSA radiologist.

Figure 2. Block diagram of the bilateral CAD system for mass detection on mammograms.

Figure 3. An example of performing the mass candidate identification. (a) an original mammogram, (b) the detected breast boundary of (a), a mass is marked by the arrow, (c) the detected mass candidates of (a).

Figure 4. An example of obtaining the corresponding ROI of a mass candidate on the contralateral mammogram. (a) mass candidate on the left MLO view at *m*, (b) corresponding ROI on the right MLO view at *m'*.

Figure 5. An example of obtaining the corresponding ROI based on the modified regional registration technique. (a) the nipple location (*o*), the shifted origin (*n*), and the mass candidate (*m*), (b) corresponding ROI on the contralateral mammogram.

Figure 6. (a) Mammogram containing a mass marked by the rectangular box. (b) A contralateral mammogram of (a) and the rectangular box is the corresponding ROI of the mass in (a) estimated by the automated regional registration technique. (c) ROI extracted from (a) containing a mass detected at the prescreening stage but excluded at the final stage of the unilateral CAD system. (d)

The corresponding ROI in the contralateral breast. Bilateral analysis of this ROI pair increased the likelihood score of the mass which was then detected as a TP in the bilateral CAD system.

Figure 7. (a) Mammogram and the rectangular ROI containing a mass candidate. (b) The contralateral mammogram of (a) and the rectangular box is the corresponding ROI of the mass candidate in (a). (c) ROI extracted from (a) containing normal tissue detected at the prescreening stage and included as an FP at the final stage of the unilateral CAD system. (d) The corresponding ROI in the contralateral breast. Bilateral analysis of this ROI pair reduced the likelihood score of the normal tissue which then became a TN in the bilateral CAD system.

Figure 8. (a) Image-based and (b) case-based average test FROC curves from the unilateral and the bilateral CAD systems. The FP rates were estimated from detection on mammograms in the test subsets with masses.

Figure 9. (a) Image-based and (b) case-based average test FROC curves from the unilateral and the bilateral CAD systems. The FP rates were estimated from detection on mammograms in the no-mass data set.

Figure 10. (a) Image-based and (b) case-based average test FROC curves from the unilateral and bilateral CAD systems for detection on cases with malignant masses only. The FP rates were estimated from in the same data set.

Figure 11. (a) Image-based and (b) case-based average test FROC curves from the unilateral and bilateral CAD systems for detection on cases with malignant masses only. The FP rates were estimated from the no-mass data set.

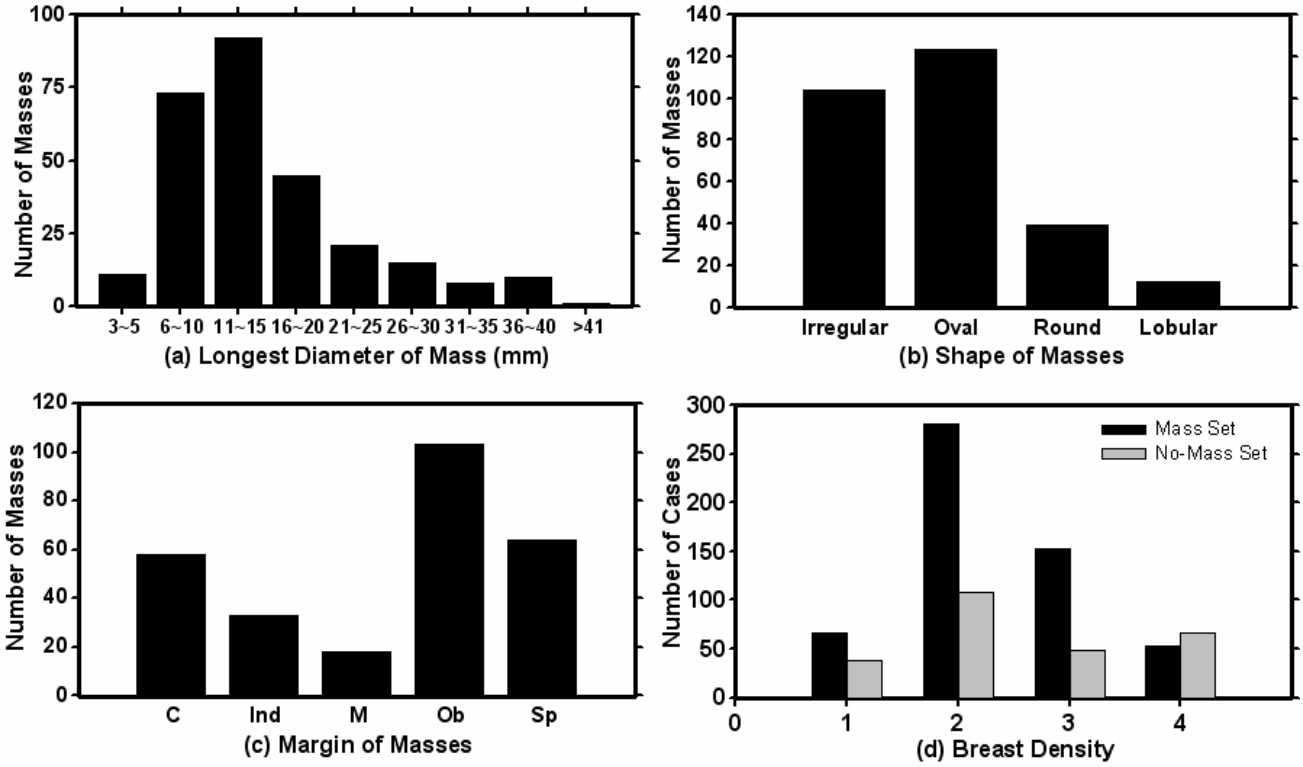


Figure 1. The characteristics of our mass data set. (a) distribution of mass sizes, (b) distribution of mass shapes, (c) distribution of mass margins, C: circumscribed, Ind: indistinct, M: microlobulated, Ob: obscured, Sp: spiculated, (d) distribution of the breast density in terms of BI-RADS category
630 estimated by an MQSA radiologist.

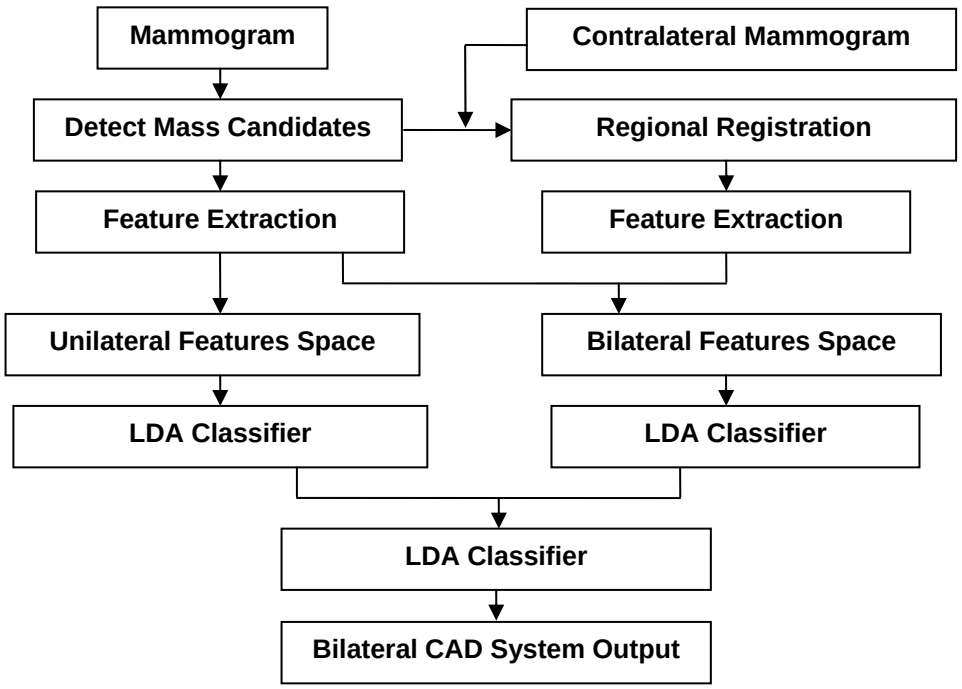


Figure 2. Block diagram of the bilateral CAD system for mass detection on mammograms.

635

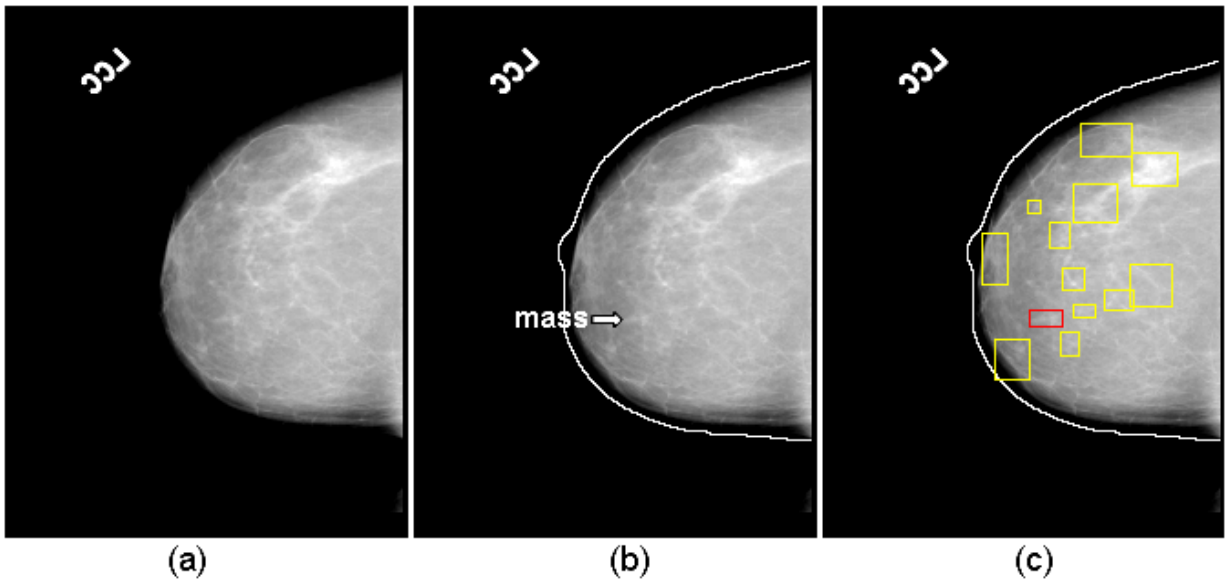


Figure 3. An example of performing the mass candidate identification. (a) an original mammogram, (b) the detected breast boundary of (a), a mass is marked by the arrow, (c) the detected mass candidates of (a).

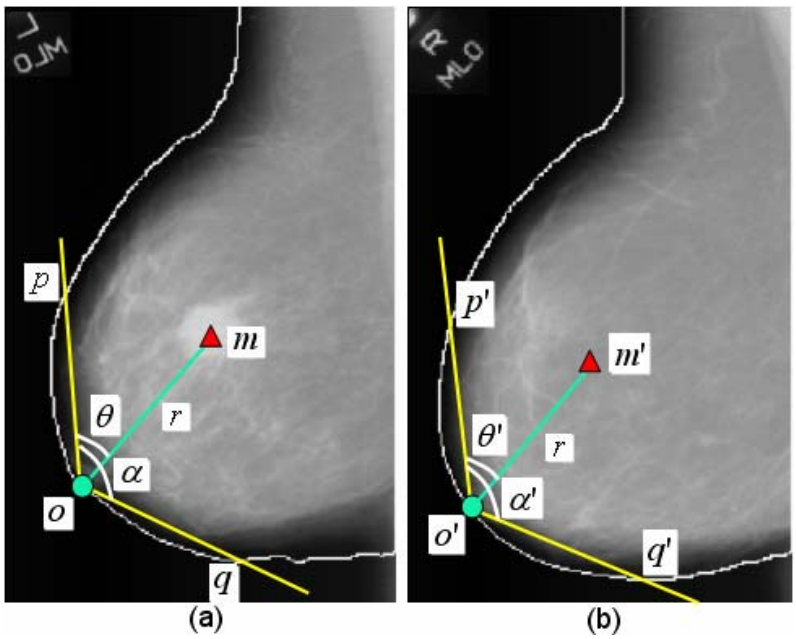


Figure 4. An example of obtaining the corresponding ROI of a mass candidate on the contralateral mammogram. (a) mass candidate on the left MLO view at *m*, (b) corresponding ROI on the right MLO view at *m'*.

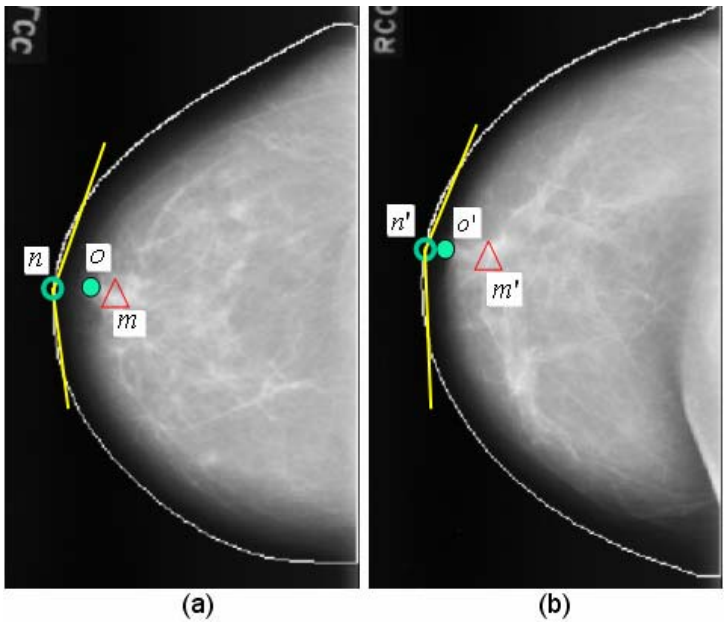
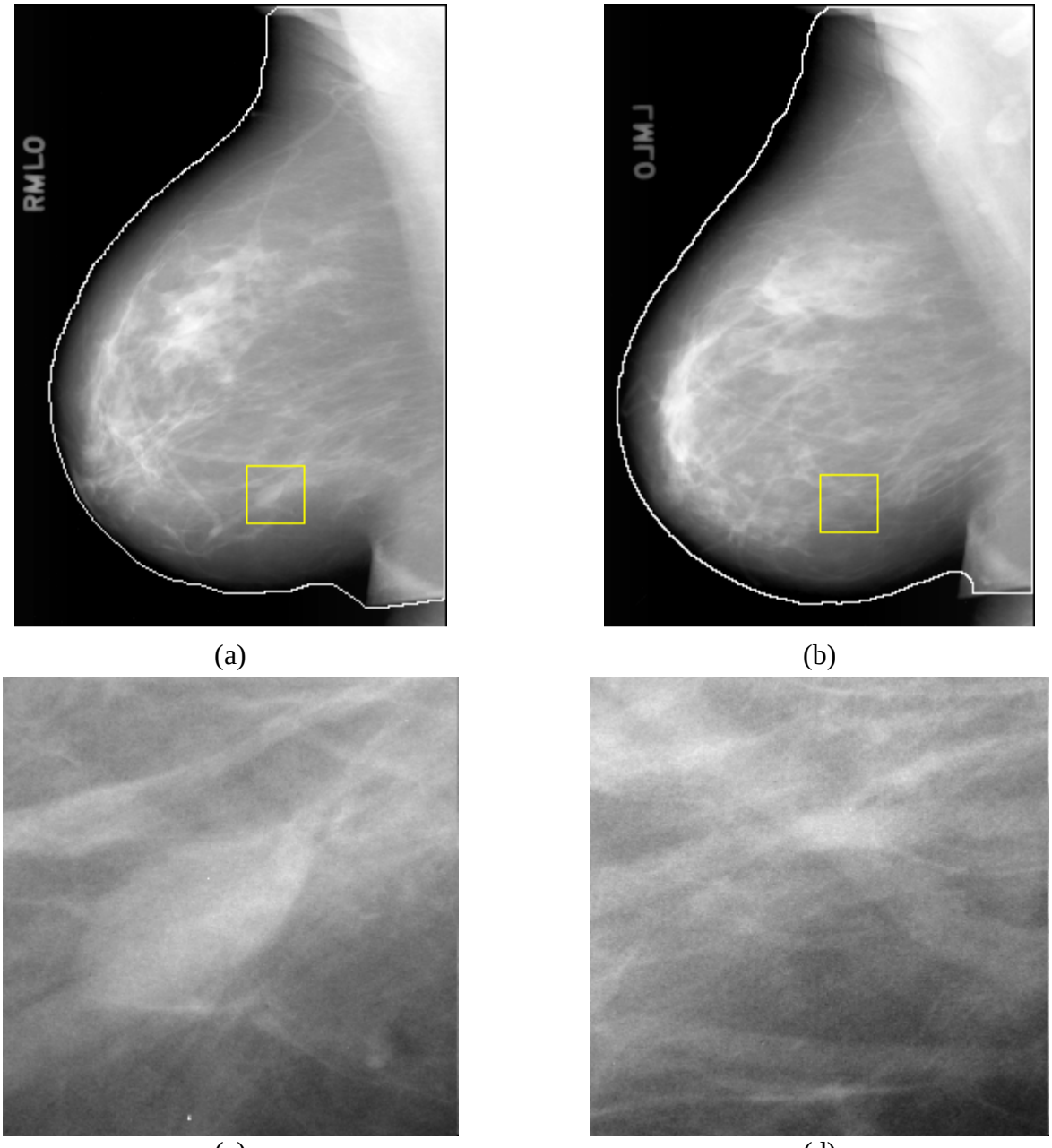


Figure 5. An example of obtaining the corresponding ROI based on the modified regional registration technique. (a) the nipple location (*o*), the shifted origin (*n*), and the mass candidate (*m*), (b) corresponding ROI on the contralateral mammogram.

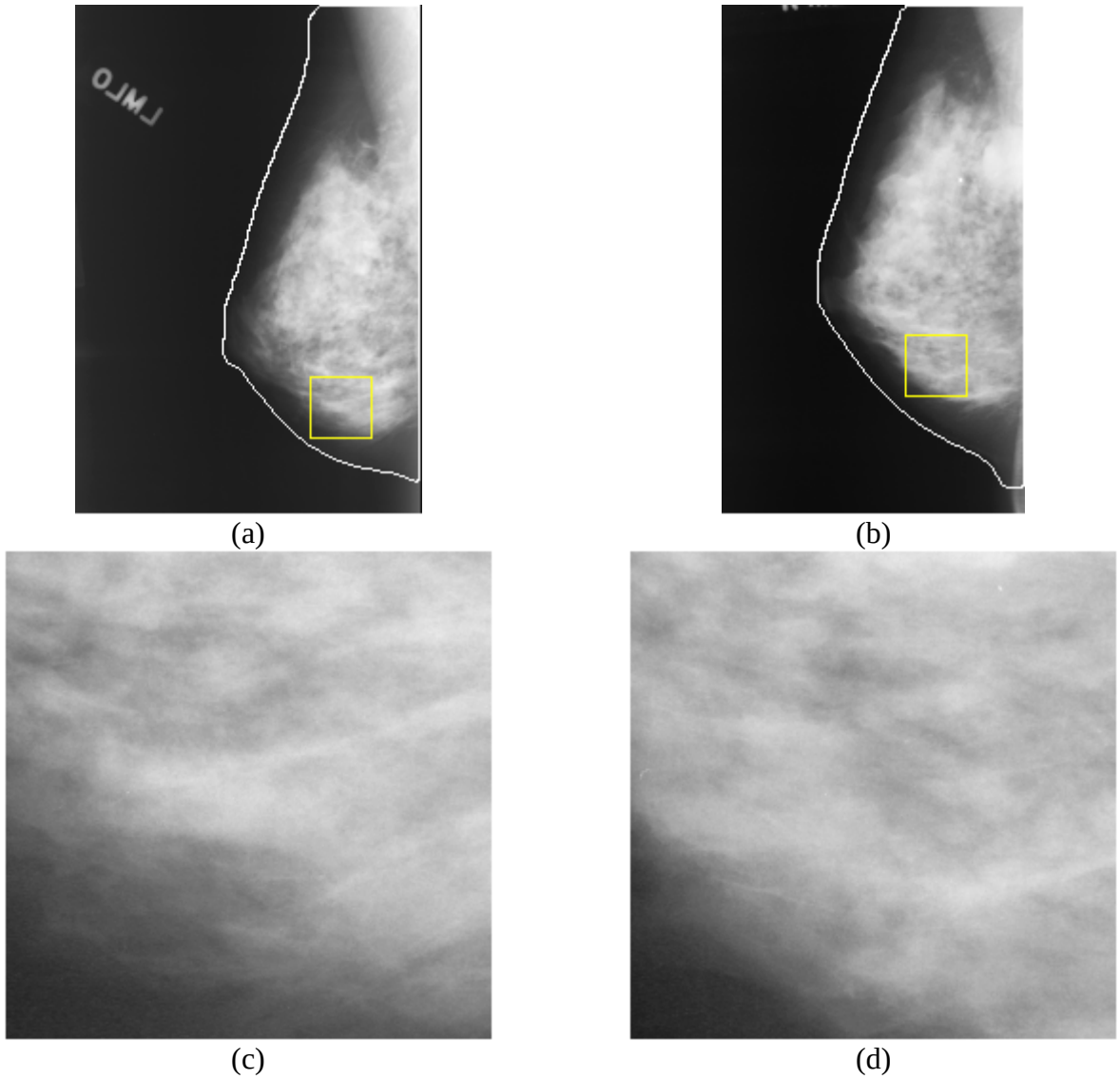


(a) (b)

(c) (d)

Figure 6. (a) Mammogram containing a mass marked by the rectangular box. (b) A contralateral mammogram of (a) and the rectangular box is the corresponding ROI of the mass in (a) estimated by the automated regional registration technique. (c) ROI extracted from (a) containing a mass detected at the prescreening stage but excluded at the final stage of the unilateral CAD system. (d) The corresponding ROI in the contralateral breast. Bilateral analysis of this ROI pair increased the likelihood score of the mass which was then detected as a TP in the bilateral CAD system.

670



(a) (b)

(c) (d)

Figure 7. (a) Mammogram and the rectangular ROI containing a mass candidate. (b) The contralateral mammogram of (a) and the rectangular box is the corresponding ROI of the mass candidate in (a). (c) ROI extracted from (a) containing normal tissue detected at the prescreening stage and included as an FP at the final stage of the unilateral CAD system. (d) The corresponding ROI in the contralateral breast. Bilateral analysis of this ROI pair reduced the likelihood score of the normal tissue which then became a TN in the bilateral CAD system.

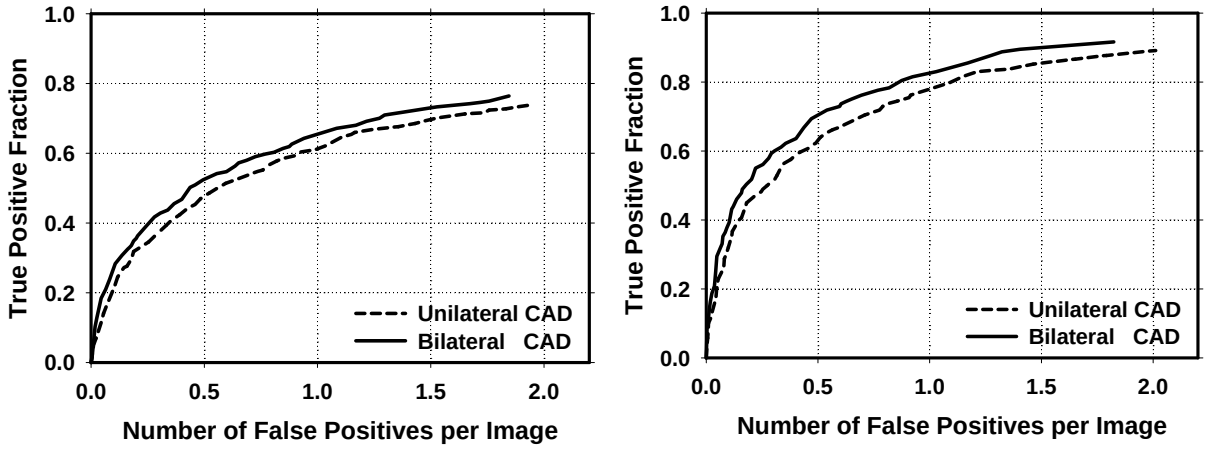


Figure 8. (a) Image-based and (b) case-based average test FROC curves from the unilateral and the bilateral CAD systems. The FP rates were estimated from detection on mammograms in the test subsets with masses.

675

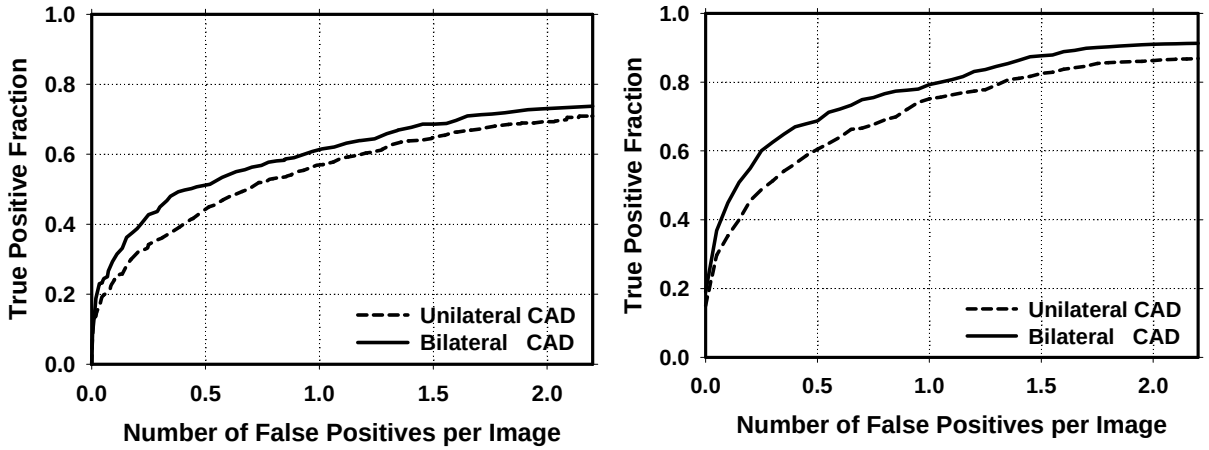
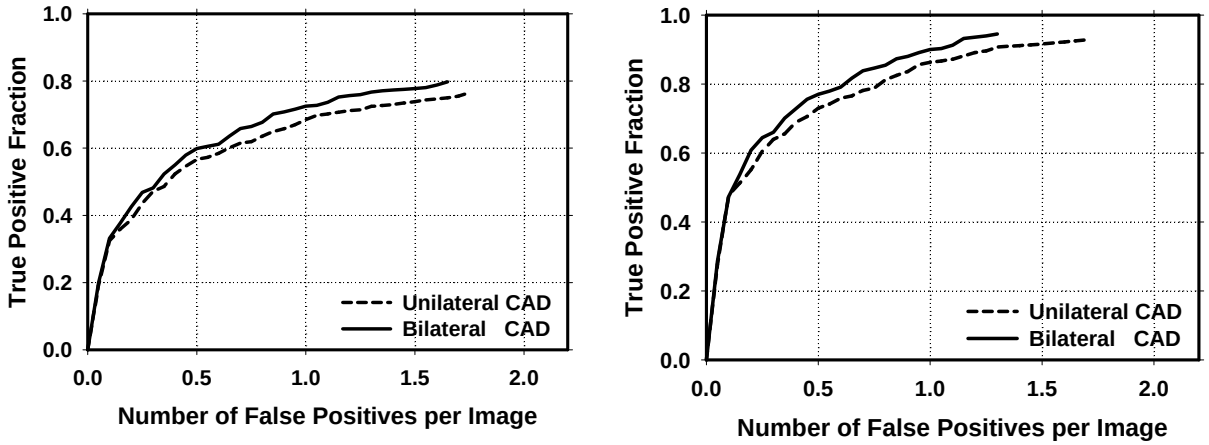


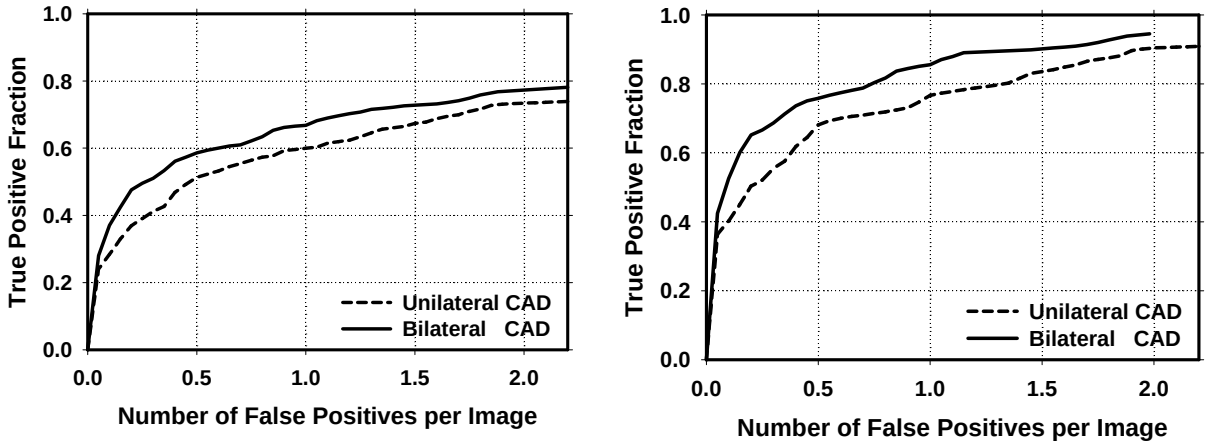
Figure 9. (a) Image-based and (b) case-based average test FROC curves from the unilateral and the bilateral CAD systems. The FP rates were estimated from detection on mammograms in the no-mass data set.

680



(a) (b)

Figure 10. (a) Image-based and (b) case-based average test FROC curves from the unilateral and bilateral CAD systems for detection on cases with malignant masses only. The FP rates were estimated from in the same data set.



(a) (b)

Figure 11. (a) Image-based and (b) case-based average test FROC curves from the unilateral and bilateral CAD systems for detection on cases with malignant masses only. The FP rates were estimated from the no-mass data set.

685

Computer aided detection of breast masses on prior mammograms

Jun Wei*, Berkman Sahiner, Heang-Ping Chan, Lubomir M. Hadjiiski, Marilyn A. Roubidoux, Mark A. Helvie, Jun Ge, Chuan Zhou, Yi-Ta Wu
Department of Radiology, The University of Michigan, Ann Arbor, MI 48109

ABSTRACT

An important purpose of a CAD system is that it can serve as a second reader to alert radiologists to subtle cancers that may be overlooked. In this study, we are developing new computer vision techniques to improve the detection performance for subtle masses on prior mammograms. A data set of 159 patients containing 318 current mammograms and 402 prior mammograms was collected. A new technique combining gradient field analysis with Hessian analysis was developed to prescreen for mass candidates. A suspicious structure in each identified location was initially segmented by seed-based region growing and then refined by using an active contour method. Morphological, gray level histogram and run-length statistics features were extracted. Rule-based and LDA classifiers were trained to differentiate masses from normal tissues. We randomly divided the data set into two independent sets; one set of 78 cases for training and the other set of 81 cases for testing. With our previous CAD system, the case-based sensitivities on prior mammograms were 63%, 48% and 32% at 2, 1 and 0.5 FPs/image, respectively. With the new CAD system, the case-based sensitivities were improved to 74%, 56% and 35%, respectively, at the same FP rates. The difference in the FROC curves was statistically significant ($p<0.05$ by AFROC analysis). The performances of the two systems for detection of masses on current mammograms were comparable. The results indicated that the new CAD system can improve the detection performance for subtle masses without a trade-off in detection of average masses.

Keywords: computer-aided detection, prior mammogram, mass detection, AFROC analysis

1. INTRODUCTION

Breast cancer is one of the leading causes of cancer mortality among women[1]. Studies indicate that radiologists do not detect all carcinomas that are visible upon retrospective analyses of the images[2-8]. Computer-aided diagnosis (CAD) is considered to be one of the promising approaches that may improve the sensitivity of mammography[9, 10].

An important application of a CAD system is to serve as a second reader to alert radiologists to subtle cancers that may be overlooked. Masses retrospectively seen on prior mammograms represent the difficult cases that are more likely to be missed by radiologists. To study the ability of a CAD system in detecting subtle cancers, one way is to evaluate its accuracy in detecting missed cancers on prior mammograms. Our previous experiences indicate that CAD schemes trained with cancers on current images do not perform well in detecting masses seen retrospectively on prior images[11]. In this study, we designed new techniques to improve the detection performance for subtle masses on prior mammograms and also evaluated the new CAD system on both prior and current mammograms by comparing with our previously developed CAD system[12].

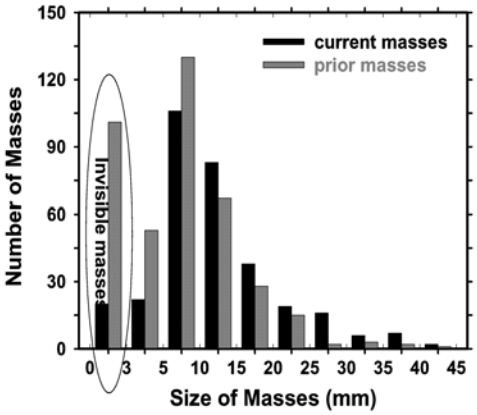
2. MATERIALS AND METHODS

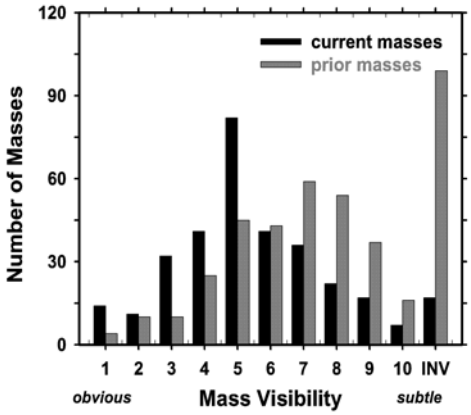
2.1 Materials

All mammograms in this study were collected from patient files in the Department of Radiology at the University of Michigan with Institutional Review Board (IRB) approval. The mammograms were digitized with a LUMISYS 85 laser film scanner with a pixel size of 50μm×50μm and 4096 gray levels. The scanner was calibrated to have a linear relationship between gray levels and optical densities (O.D.) from 0.1 to greater than 3 O.D. units. The nominal O.D.

* jvwei@umich.edu, phone: 734-647-8553, CGC B2103, 1500 E. Medical Center Dr., Ann Arbor, MI 48109-0904

Medical Imaging 2007: Computer-Aided Diagnosis, edited by Maryellen L. Giger, Nico Karssemeijer,
Proc. of SPIE Vol. 6514, 651405, (2007) · 1605-7422/07/$18 · doi: 10.1117/12.713768


150
120
90
60
30
0
Number of Masses
current masses
prior masses
invisible masses
0 3 5 10 15 20 25 30 35 40 45
Size of Masses (mm)


120
90
60
30
0
Number of Masses
current masses
prior masses
1 2 3 4 5 6 7 8 9 10 INV
obvious
Mass Visibility
subtle

2.2 Methods

2.2.1 CAD System Overview

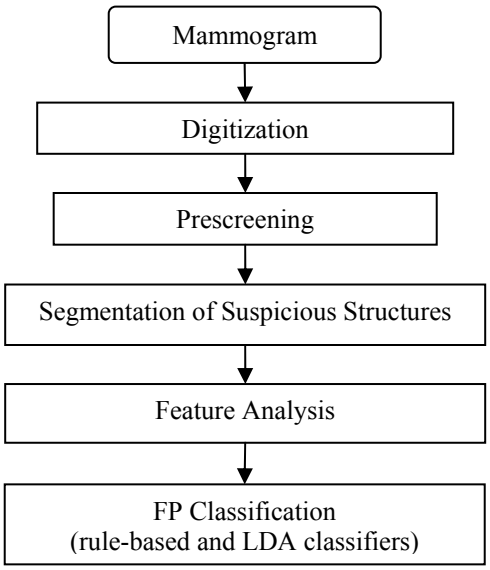


Figure 3. Block diagram of a single CAD system for mass detection on mammograms.

Our CAD system consists of five processing steps: 1) pre-screening of mass candidates, 2) identification of suspicious objects, 3) extraction of morphological and texture features, and 4) classification between the normal and the abnormal regions by using rule-based and LDA classifiers. The block diagram for the CAD system is shown in Figure 3.

For the pre-screening stage, we developed a new prescreening technique in which gradient field analysis was combined with Hessian analysis to identify mass candidates. Both gradient field and Hessian analyses were designed to enhance circular structures on mammograms and to suppress the objects with other shapes. Gradient field analysis used the information of gradient field directions and Hessian analysis used the second derivatives by solving for the eigenvalues of the Hessian matrix. After this enhancement filtering, the local maxima within the breast region were identified as the mass candidates on each mammogram. The suspicious structure in each identified location was initially extracted by a seed-based region growing method. An active contour method was then used to further refine the initial segmentation. Morphological, gray level histogram and run-length statistics (RLS) features were extracted from the original region of interest (ROI) and the orientation field of the ROI for reduction of FPs.

2.2.2 Training and test CAD system

The hold-out method was used for training and testing our CAD system. We randomly separated the entire data set by case into two independent subsets, the training subset including 78 cases with 156 current and 200 prior mammograms and the test subset including 81 cases with 162 current and 202 prior mammograms. The training included selection of proper parameters and features for the classifier in the CAD system. Once the training was completed, the parameters and features were fixed for testing. The new system was trained by using prior mammograms in the training set only. The performance of the new system was compared with that of the previous CAD system on the current and prior mammograms in the test set.

During training, feature selection with stepwise LDA was employed to obtain the best feature subset and reduce the dimensionality of the feature space to design an effective classifier. The detailed procedure has been described

elsewhere[13]. Briefly, at each step one feature was entered or removed from the feature pool by analyzing its effect on the selection criterion, which was chosen to be the Wilks' lambda in this study. Since the appropriate threshold values for feature entry, feature elimination, and tolerance of feature correlation were unknown, we used an automated simplex optimization method to search for the best combination of thresholds in the parameter space. The simplex algorithm used a leave-one-case-out resampling method within the training subset to select features and estimate the weights for the LDA classifier. To have a figure-of-merit to guide feature selection, the test discriminant scores from the left-out cases were analyzed using receiver operating characteristic (ROC) methodology. The accuracy for classification of masses and FPs was evaluated as the area under the ROC curve, A_z. In this approach, feature selection was performed without the left-out case so that the test performance would be less optimistically biased. However, the selected feature set in each leave-one-case-out cycle could be slightly different because every cycle had one training case different from the other cycles. In order to obtain a single trained classifier to apply to the hold-out test subset, a final stepwise feature selection was performed with the best combination of thresholds, found in the simplex optimization procedure, on the entire training subset to obtain the final set of features and estimate the weights of the LDA. Note that the entire process of feature selection and classifier weight estimation was performed within the training subset. The LDA classifier with the selected feature set was then fixed and applied to the test subset.

2.2.3 Evaluation methods

We used a free-response receiver operating characteristic (FROC) method to assess the overall performance of the CAD scheme on this image set. An FROC curve was obtained by plotting the mass detection sensitivity as a function of FP marks per image as the decision threshold on the LDA classifier scores varied. The detected individual objects were compared with the "true" mass locations marked by the experienced radiologist, as described above. A detected object was labeled as TP if the overlap between the bounding box of the detected object and the bounding box of the true mass relative to the larger of the two bounding boxes was over 25%. Otherwise, it would be labeled as FP. The 25% threshold was selected as described in our previous study[14].

FROC curves were presented on a per-image and a per-case basis. For image-based FROC analysis, the mass on each mammogram was considered an independent true object; the sensitivity was thus calculated relative to the number of visible masses by image, which was 149 and 151, respectively, for the current and prior test subset. For case-based FROC analysis, the same mass imaged on the two-view mammograms was considered to be one true object and detection of either or both masses on the two views was considered to be a TP detection; the sensitivity was thus calculated relative to the number of masses by case, which was 81 and 90, respectively, for the current and prior test subset. The test FROC curve for a given mass subset was estimated by counting the detected masses on the test mass subset for the sensitivity. The FP marker rate was estimated from FPs detected in the same test subset. The average number of FP marks per image produced by the CAD system at a given sensitivity was estimated by counting the detected objects in these cases at the corresponding decision threshold.

In order to compare the performance of our CAD systems statistically, we employed the alternative free-response ROC (AFROC) method[15]. In the AFROC method, the FROC data are first transformed by counting the number of false-positive images (FPI) instead of the FPs per image. The LDA score of an FPI is determined by the FP object with the highest score on the image regardless of how many lower scores FP objects are made on the same image. The ROCKIT curve fitting software and statistical significance tests for ROC analysis developed by Metz et al. [16] can then be used to analyze the AFROC data.

3. EXPERIMENTAL RESULTS

Figures 4 and 5 showed the image-based and case-based FROC curves for detection of masses on prior mammograms, respectively. The case-based sensitivities for detection of masses on the prior mammograms (typically subtle masses) in the test subset were 56%, and 35% at 1 and 0.5 FPs/image by using the new CAD system in comparison to 48%, and 32% at the same FP rates by using the previous CAD system. The improvement with the new system on prior mammograms was statistically significant ($p = 0.036$). When the new system was applied to the detection of masses

on the current mammograms (typically average masses) in the test subset, the case-based sensitivities were 77% and 70% at 1 and 0.5 FPs/image in comparison to 75% and 56% at the same FP rates by using the previous CAD system. The difference in the two FROC curves for detection of average masses on current mammograms was not statistically different ($p = 0.184$). Image-based and case-based FROC curves for detection of masses on current mammograms were shown in Figures 6 and 7, respectively.

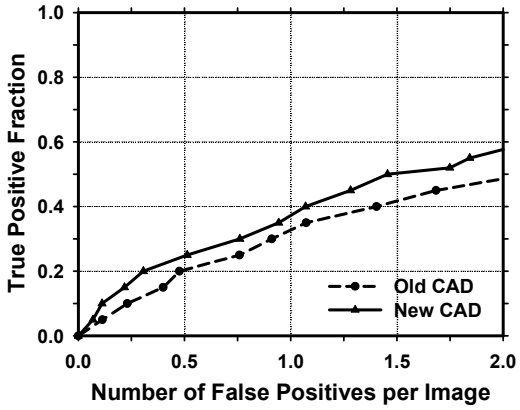


Figure 4. Image-based test FROC curves on prior mammograms. Old CAD: detection by the previous CAD system trained on both current and prior mammograms. New CAD: detection by the CAD system trained on prior mammograms.

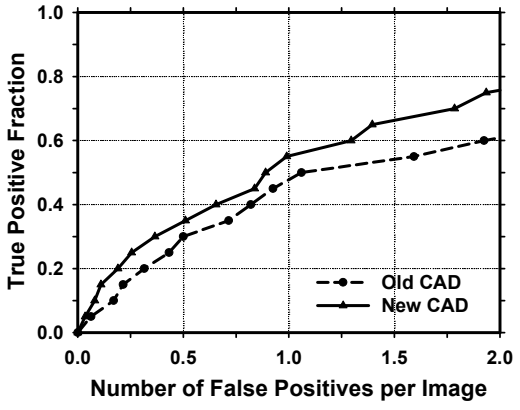


Figure 5. Case-based test FROC curves on prior mammograms. Old CAD: detection by the previous CAD system trained on both current and prior mammograms. New CAD: detection by the CAD system trained on prior mammograms.

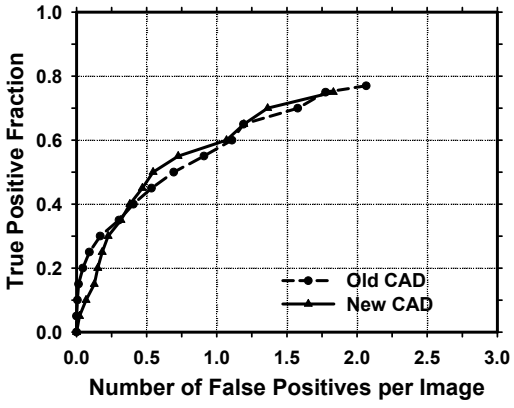


Figure 6. Image-based test FROC curves on current mammograms. Old CAD: detection by the previous CAD system trained on both current and prior mammograms. New CAD: detection by the CAD system trained on prior mammograms.

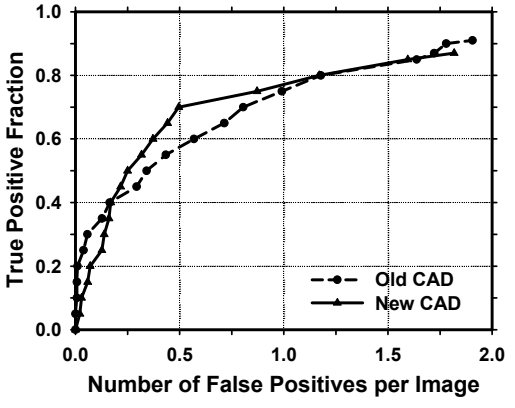


Figure 7. Case-based test FROC curves on current mammograms. Old CAD: detection by the previous CAD system trained on both current and prior mammograms. New CAD: detection by the CAD system trained on prior mammograms.

Table 1. Estimation of the statistical significance in the difference between the FROC performances of the previous CAD system trained on both current and prior mammograms and the proposed CAD system trained on prior mammograms.

	A_1 (AFROC)	
	Current Test Set	Prior Test Set
Old CAD	0.51	0.26
New CAD	0.50	0.31
p-value	0.184	0.036

4. DISCUSSION AND CONCLUSIONS

In this study, we improved the accuracy of a CAD system for detection of subtle masses on prior mammograms. A new prescreening method was developed to improve the sensitivity of mass detection. A new mass segmentation method that combined a seed-based region growing method with active contour method was also designed. RLS features were extracted from the original ROIs and the newly derived orientation field of the ROIs for FPs reduction. Our CAD system can significantly improve the performance of mass detection on prior mammograms without a trade-off in the detection of masses on current mammograms. It is expected that the new CAD system can increase the overall accuracy for detection of subtle early-stage breast cancers.

ACKNOWLEGMENTS

This work is supported by USPHS grant CA95153, U. S. Army Medical Research and Materiel Command grants DAMD 17-02-1-0214 and W81XWH-04-1-0475. The content of this paper does not necessarily reflect the position of the funding agencies and no official endorsement of any equipment and product of any companies mentioned should be inferred. The authors are grateful to Charles E. Metz, Ph.D., for the LABROC and ROCKIT programs.

REFERENCES

1. "Cancer statistics 2006 presentation, www.Cancer.Org," American Cancer Society (2006).
2. B. J. Hillman, L. L. Fajardo, T. B. Hunter, B. Mockbee, C. E. Cook, R. M. Hagaman, J. C. Bjelland, C. S. Frey and C. J. Harris, "Mammogram interpretation by physician assistants," *AJR* **149**, 907-911, 1987.
3. L. W. Bassett, D. H. Bunnell, R. Jahanshahi, R. H. Gold, R. D. Arndt and J. Linsman, "Breast cancer detection: One versus two views," *Radiology* **165**, 95-97, 1987.
4. M. G. Wallis, M. T. Walsh and J. R. Lee, "A review of false negative mammography in a symptomatic population," *Clinical Radiology* **44**, 13-15, 1991.
5. J. A. Harvey, L. L. Fajardo and C. A. Innis, "Previous mammograms in patients with impalpable breast carcinomas: Retrospective vs blinded interpretation," *AJR* **161**, 1167-1172, 1993.
6. R. E. Bird, T. W. Wallace and B. C. Yankaskas, "Analysis of cancers missed at screening mammography," *Radiology* **184**, 613-617, 1992.
7. C. A. Beam, P. M. Layde and D. C. Sullivan, "Variability in the interpretation of screening mammograms by US radiologists - findings from a national sample," *Archives of Internal Medicine* **156**, 209-213, 1996.
8. V. Beam, D. Sullivan and P. Layde, "Effect of human variability on independent double reading in screening mammography," *Academic Radiology* **3**, 891-897, 1996.
9. F. Shtern, C. Stelling, B. Goldberg and R. Hawkins, "Novel technologies in breast imaging: National cancer institute perspective," *2nd Postgraduate Course Syllabus*. 153-156, 1995.
10. C. J. Vyborny, "Can computers help radiologists read mammograms?," *Radiology* **191**, 315-317, 1994.
11. J. Wei, H.-P. Chan, B. Sahiner, L. M. Hadjiiski, M. A. Helvie, M. A. Roubidoux, C. Zhou and J. Ge, "Dual system approach to computer-aided detection of breast masses on mammograms," *Medical Physics* **33**, 4157-4168, 2006.
12. J. Wei, B. Sahiner, L. M. Hadjiiski, H.-P. Chan, M. A. Helvie, M. A. Roubidoux, C. Zhou, J. Ge and Y. Zhang, "Two-view information fusion for improvement of computer-aided detection (CAD) of breast masses on mammograms," *Proc. SPIE* **6144**, 241-247, 2006.

13. J. Wei, B. Sahiner, L. M. Hadjiiski, H. P. Chan, N. Petrick, M. A. Helvie, M. A. Roubidoux, J. Ge and C. Zhou, "Computer aided detection of breast masses on full field digital mammograms," *Medical Physics* **32**, 2827-2838, 2005.
14. N. Petrick, H. P. Chan, B. Sahiner, M. A. Helvie, S. Paquerault and L. M. Hadjiiski, "Breast cancer detection: Evaluation of a mass detection algorithm for computer-aided diagnosis: Experience in 263 patients.," *Radiology* **224**, 217-224, 2002.
15. D. P. Chakraborty and L. H. L. Winter, "Free-response methodology: Alternate analysis and a new observer-performance experiment," *Radiology* **174**, 873-881, 1990.
16. C. E. Metz, "ROC methodology in radiologic imaging," *Investigative Radiology* **21**, 720-733, 1986.